Quantitative Analysis of Pathological Strips using Machine Learning

Author

Babar Zeb

00000172350

MS-16(CSE)

Thesis Supervisor

Dr. Aimal Khan

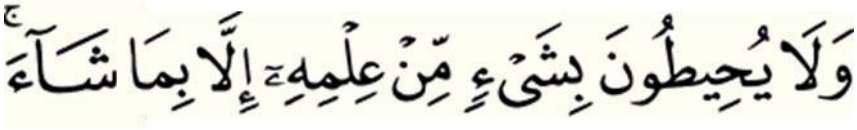
Department of Computer and Software Engineering,

College of Electrical and Mechanical Engineering,

National University of Sciences and Technology, Islamabad

July 2020

In the name of God, most Gracious, most Compassionate



And they can’t encompass anything from His knowledge, but to extend He wills [2:255]

Quantitative Analysis of Pathological Strips using Machine Learning

Author
Babar Zeb
00000172350

A thesis submitted in partial fulfillment of the requirements for the degree of
MS Computer Software Engineering

Thesis Supervisor:
DR. Aimal Khan

_____

DEPARTMENT OF COMPUTER AND SOFTWARE ENGINEERING, COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING, NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY,
ISLAMABAD
July 2020

Language Correctness Certificate

This thesis has been read by an expert of the English language and is found free of typing, syntax, semantic, grammatical, and spelling mistakes. The stated is also according to the format given by the university.

Signature of Student
Babar Zeb
Registration Number
00000172350

Signature of Supervisor
Dr. Aimal Khan

Declaration

I certify that this research work titled "*Quantitative Analysis of Pathological Strips using Machine Learning"* is my work. The work has not been presented elsewhere for assessment. Material derived from other sources has been properly acknowledged/referred.

Signature of Student
Babar Zeb
MS-16(CSE)
00000172350

Copyright Statement

This page is intentionally left blank

Acknowledgments

I am extremely thankful to my ALLAH, The Almighty, for bestowing countless blessings upon me and vesting in my patience and ability to complete this manuscript. It is merely because of His kindness and assistance which enabled me to endure and succeed. I am very grateful to my supervisor **Dr. Aimal Khan** for his guidance and efforts. Throughout this journey, he has always been there to guide me and boost my confidence. He has been a source of great knowledge and his expertise made this possible. I want to thank my teacher, **Dr. Farhan Hussain**, for his time, and ideas. Furthermore, I want to extend my heartiest gratitude to **Dr. Aarsalan Shaukat** and **Dr. Wasi Haider Butt** for their availability, guidance, and support. I want to appreciate my Father, who left no stone unturned to educate me. I want to thank my mother for her prayers and love. I want to thank **Ma'am Mahvesh Ali Khan** and Brother-in-law **Ateeq Muhammad khan** for their financial support. I am grateful to my brothers, and friends **Younas Khan**, **Muhammad Faisal Masood**, **Muhammad Asad**, **Iqra Tahir,** and **Uzair Rasheed** for motivating and encouraging me.

Dedicated to my Daji, Bibi, and better half

Abstract

Urinalysis is a significant technique used for inspecting the urinary system. Urine has numerous chemical materials secreted; these materials can be used to diagnose diseases and conditions such as urinary tract infection, diabetes, kidney diseases, and pregnancy, at the earliest. These diseases affect people of every age, gender, and profession. If not diagnosed timely it can cause death in a matter of months if not days. Mostly, urinalysis is used for diagnosing it, but studies show that it is a costly procedure and is prone to human error. Moreover, numerous studies and research works have been carried out to present techniques for diagnosing these diseases, and they have achieved success to a great extent. However, there is still room for improvement, and until 100% accurate results are achieved, the struggle for diagnosing and treating diseases would continue. This thesis has conducted a systematic literature review to thoroughly analyze the existing literature and look for gaps in it. 33 research studies that have been published during 2007-2019 related to the specified domain have been identified and analyzed. This leads to the identification of 10 methods, 8 technologies, 12 challenges, and 10 diseases. The main objective of this thesis is to present a technique that can be used to develop an automated system for urinalysis. The proposed technique has been able to outclass the previously presented techniques in terms of accuracy. It has been able to achieve the highest accuracy because of the different experimentations carried out to analyze their results. Besides this, the results of the proposed methodology have also been evaluated and compared with one of the previously used methods i.e. Euclidean Distance. The dataset used in this manuscript contains the urine test strip RGB images, which is developed using MATLAB. Different noises i.e. Gaussian and salt & pepper have been applied on the dataset (images) to increase the number of instances. Initially, the dataset was comprised of three regent color strip images, which are used as inputs for Euclidean distance. After this, segmentation has been applied to the images and then the segmented images have been used for experimentation purposes. The accuracy vs noise variance graphs for the models have been presented in **Figure 5-10**. The experimentation results reveal that the regression model with the CNN classifier has a higher accuracy as compared to Euclidean Distance.

Table of Contents

List of Equations

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

Daily newspapers bring the news of technological advancement, advancement in machines, and automation, computerized systems are being developed to provide ease and precision in human effort and work. Almost every field has been automated which includes the medical field as well where the automated systems need to be faultless because it is related to human life, which means the automated system should i.e. accurate, efficient, robust, etc. The biggest advantage of an automated system is its speed. In manual systems, things are time-consuming and domain expert's time is precious; that is why an automated system provides them ease, reduces their effort, and saves their time. The system automation consists of computers and modern tools and the internet. Think of a manual system, they have a register to store the information of sales which is a slow process and requires a lot of effort; the same process can be done with Point of Sale which makes all the things smooth, easy, and precise. The PoS saves the records and details of the inventory, details of the sales process, and other records can be accessed with the least amount of effort. Similarly, in the medical field manual diagnosis of diseases is a very difficult task; it's a slow and costly process.

The automation of the healthcare system has been emphasized for the last 30 years or so. Automated systems have been successfully deployed in hospitals to assist the doctors and their technical staff successfully, but there is always room for improvement in the automation of the healthcare systems. No mistake can be tolerated because a minor mistake in sensitive information can be life-threatening for patients.

The other things include unstructured medical records, but machine learning can handle it with some data pre-processing methods. The other constraint is real-time systems should have a real-time response; these automated systems cannot afford to be slow because a mistake can lead to loss of human life. These time constraints can be achieved via tackling the redundant features by feature selection techniques

1.1Overview

With time diseases i.e. Kidney disease, Urinary Tract Infection, Diabetes and Tumor have risked and taken more lives. The problem is not limited to gender, age, or any other demographic characteristics. Therefore, the accurate prediction, preferably, at an early stage is extremely vital to save someone's life. Urinalysis has been widely used for some time in disease diagnosis and the current condition of patients. This process is like the normal medical laboratory e.g. tests via chemical test strips, urine flow cytometry, and microscopy [1]. Mostly urinalysis is used for finding the chemical configuration in urine, conducted using urine dipsticks reagents. Normally a urine stick is immersed in the urine sample which gives out reactive color on the strip; that color is compared with the color chart of dipstick [2]. Some functions of the body can be determined with the process of urinalysis. This process has a high usage worldwide and has significance in diagnoses of diseases such as urinary tract infection, renal diseases, bladder inflammation, diabetes, and glucosuria but this process is slow and expensive while in comparison to this process the other methods available are fast but not much reliable [3]. In the literature on machine learning, a stupendous amount of work has already been done to diagnose diseases related o urinalysis. However, what is mind-boggling is that an agreement of which technique is best for which kind of data or problem is yet to be established. But what is agreed upon is that feature selection, noise reduction, and techniques of optimization help enhance the accuracy and efficiency of the existing techniques. Besides this, it is also necessary to make the system available for the people so they can assess their health daily. For this purpose, we are going to develop a mobile application in the future to save time and money. These are the reasons why this research study has been conducted.

1.2Background and Motivation

Kidney disease, Urinary Tract Infection, Diabetes, and Tumor are the most fatal diseases and have been causing death all around the world. It affects people of every age, gender, and profession, and many factors are responsible to cause the above-mentioned diseases. If not diagnosed timely it can cause death in a matter of months if not days. Urinalysis is used most for diagnosing it, but studies show that it has side effects and is costly as well. Moreover, numerous studies and research works have been carried out to diagnose these diseases and to a great extent,

scientists have achieved success. However, there is still room for improvement, and until 100% of results are achieved the struggle for diagnosing and treating diseases would continue.

The literature has a huge amount of work on pathological strips and urinalysis for healthcare. There also exists a decent amount of work on urinalysis but very few studies have tried to enhance and use machine learning techniques for urinalysis. This is the reason why they need for exploring this factor has been felt. Moreover, the fact that an automated urinalysis system can be used to assist physicians, clinicians, and other experts for saving a life is another motivating factor of this thesis.

1.3Objective and Contributions

Urinary Tract Infection, Diabetes, and Tumor have been causing deaths irrespective of age, gender, or any other demographics, around the globe. When such diseases are not diagnosed timely, consequently death becomes certain. The existing clinical techniques for instance: urinalysis is used for detecting Urinary Tract Infection, Diabetes, and Tumor, but it has reportedly side effects and it is a very expensive procedure. Modernized hospitals often contain patients' data, that data can be used by machine learning to predict or classify patients. The main objective of this thesis is to present a technique that can be used to assist physicians, clinicians, and other medical experts to diagnose the above-listed diseases, particularly, diabetes and tumor disease to save valuable lives.

A urine test strip Dataset has been used in this study as the main dataset. Different techniques are applied to the stated dataset for experimentation purposes. The final algorithm has been established, where the main contribution of this thesis is to develop a dataset, which contains RGB images of the urine test strip. Two different techniques i.e. Convolutional Neural Network and RGB to Lab Euclidean Distance has been used for evaluation. The results of both techniques are compared in chapter 5.

1.4Outline

Apart from this chapter, this manuscript has five more chapters. Chapter II is used to present the related terminology, it gives the reader some required background knowledge about pathological strips, urinalysis, machine learning, and the use of machine learning for healthcare and diagnosis

systems. Chapter II also contains a systematically conducted literature review, which is used to analyze the existing similar techniques, and gaps in the existing literature. Since numerous feature selection and weight optimization techniques have first been trialed to thrive for the best possible system, the manuscript has presented a rough methodology in Chapter III, and different experiments have been performed in the same chapter. Chapter IV proposes the final methodology and its implementation, Chapter V has the results, comparisons, and analysis and lastly, Chapter VI discusses the conclusion and future work of this thesis.

1.5 Summary

This is the first and introductory chapter of the thesis, it is used to provide the reader with an idea about the significance of information technology and particularly machine learning for pathological strips and urinalysis. It explains how technology is used for assisting and aiding physicians, clinicians, and other medical experts. What challenges are faced and how are these challenges coped with. Moreover, an overview of the thesis, its background and motivation, its objective and contributions, and its outline have also been shown in this chapter of introductory.

CHAPTER 2

LITERATURE REVIEW

2.1Introduction

Getting an insight into previous and related literature is a compulsory feature of any project be that academic or industrial. What makes a literature review effective is the fact that it should create a solid base to advance knowledge. It plays a key role in the facilitation of the development of theory, shuts areas, where an excess of research presides and explores areas where there is a need for research [1].

However, studies have proven that the conventional literature review has issues, and more often than not the results attained might be misleading, and one of the foremost tools that are used for supporting the evidence-based artifacts, mostly in the development of Systematic Literature Review. SLRs are made use of for aggregating the experiments from a defined range of various studies for answering a set of research questions. Reviews of the stated kinds employ some carefully defined criteria for determining which research papers are to be taken into consideration and analyzed [2].

In software engineering, the number of empirical studies has grown rapidly, and because of that systematic procedures should be placed to assess and aggregate the outcomes of research for providing an in-depth summary of a specific topic [3].

Keeping the importance of SLR in mind, this thesis has systematically implemented its literature review. It has predefined inclusion and exclusion criteria, i.e. existing studies that have been selected and analyzed must match the predefined criteria. These criteria are formed based on publishers, and some quality assessment criteria.

Before we could begin the discussion of our main systematic literature review, there are certain factors or areas which need to be illustrated, for instance: pathology and its types, pathological strips and, how pathological strips are being used for healthcare and urinalysis.

2.2Pathology

The study of disease is known as Pathology, more precisely when the study is about the abnormalities caused in the structure by the disease. Every aspect of the patient's health is taken in considering whether the procedure is based on diagnostic testing to state-of-the-art genetic technologies, to prevent disease, and to advise treatment. Pathology describes a disease by its characteristics such as pathology of disease “cancer”. Pathology is further divided in main three categories as clinical, anatomical, and molecular pathology. The invention of the microscope in the 19th century revolutionized pathology. It was due to the invention of the microscope that cells went under detailed study unprecedented. Now the focus shifted from studying organs to studying cells individually. As time passed and microscopes with improved capability developed, the research in pathology grew exponentially to the point of tissue and organ transplants [4].

2.2.1 Types Pathology

As the discussion above the three categories in pathology: the clinical pathology, anatomical pathology, and molecular pathology. These categories are further divided into subcategories to study different diseases in specific ways.

2.2.1.1 Anatomical Pathology

The anatomical features are studied in this type of pathology, features like removal of tissues from the human body, or sometimes the whole body in case of the autopsy, to know more about the disease. Anatomical pathology includes both aspects, looking into cells under a microscope and considering the organs such as the ruptured spleen. The chemical properties of cells and their immunological makers are also part of the pathological investigation. Some main subtypes of anatomical pathology are Histopathology, Cytopathology, and Surgical Pathology [5].

2.2.1.2 Clinical Pathology

Clinical pathology takes body fluids and tissues for laboratory tests to diagnose disease such as blood components are analyzed and analysis of cells for the identification and presence of microorganisms like bacteria. In general, this type of pathology is sometimes termed laboratory medicine which is further divided into three main types: Hematology, Chemical pathology, and Immunopathology [6].

2.2.1.3 Molecular Pathology

The abnormalities in cells and tissues are studied at the molecular level in molecular pathology, also called the study of tissue or organ disease in a body via examining the presence of the types of molecules in the cells. The aspects of clinical and anatomical pathology can be combined in molecular pathology. The major techniques include Polymerase chain reaction (PCR) for amplification of DNA, karyotype imaging of chromosomes, fluorescence labeling, and DNA microarrays [7].

2.3 Pathological Strips

Urinalysis is a testing procedure for the diagnosis of different diseases by testing samples of urine using a Urine strip also called the Pathological strip. Ten chemical pads per strip are the standard package. When submerged and removed from a urine sample a chemical reaction occurs in the pads. The timespan of soaking pads in the urine sample ranges from one to two minutes but for various tests, a larger span is also recommended. The primary step of urinalysis is a time framed examination of urine with a multi-parametric strip to diagnose a disease. Urinalysis verifies the presence of Bilirubin, Glucose, Blood, Protein, PH, Leucocytes, Urobilinogen, Specific Gravity, Ketone and Nitrate in the sample. All of these are test proven to be infections by Pathogens. Urine strip is comprised of a plastic or a paper ribbon having a width of five-millimeter. The pads infused with different chemicals, when reacting with urine, alter its color, and depicts a predefined possible result. The colors delineate different attributes related to the sample. The paper strips gather the reactants to perform a reaction I.e. pH measurement. It also can show several different findings simultaneously [8].

The following 10 reagents of pathological strips are discussed below

2.3.1 Glucose

Mostly all the detached glucose in the glomerulus is regathered in the tubule of proximal convoluted provided the regular conditions. With the rise in the glucose level of blood the ability of convoluted tubule to regather glucose outstrips (commonly known as reabsorption threshold). The glucose has a threshold between 160-180mg/dl. Test samples should be collected no less than two hours after the food is eaten due to the cause that food with high sugar can influence the results showing transitory glucosuria in a healthy person. The enzymatic reactions of glucose oxidase and peroxidase cause the finding of glucose by the strips [9].

2.3.2 Bilirubin

The deterioration of hemoglobin leads to the production of bilirubin, a pigmented compound that is released after the mononuclear phagocytes system. The worn-out red blood cells are collected and transformed into particles i.e. Protoporphyrin, Protein by bilirubin. Protoporphyrin is transformed into albumin by the blood cells and then it proceeds to run in the circulatory system. The kidney cannot block out bilirubin because it has a strong bond with protein. To solve this problem, the bilirubin is infused with glucuronic acid and a water-soluble infused bilirubin is created in the liver.

Due to the immediate excretion of bilirubin from the intestine to bile, the possibility of bilirubin to emerge in urine is null. Inside intestines bilirubin is converted into urobilinogen by the bacteria. Urobilinogen then oxidizes or excrete with either feces as stercobilin or with urine as urobilin. Blockage in the biliary duct or abnormality in the kidney's functional integrity can lead to the disturbance of the transformation process which will cause the emergence of bilirubin in urine. Due to this, the infused bilirubin breaks out into the circulation which also happens in hepatitis or hepatic cirrhosis. The finding of bilirubin in urine suggests the presence of a disease in the liver. The identification of existence or inexistence of bilirubin can assist in the diagnosis of many diseases [10].

2.3.3 Ketone

The three transitional products in fatty acids' breakdown are referred to as ketones I.e. acetoacetic acid, acetone, and beta-hydroxybutyric acid. Due to the breaking down of ketones, the possibility of its presence in urine is reduced to the lowest, as well as water, carbon dioxide, and water is produced. With the disturbance of carbohydrate metabolism, a disproportion in the metabolism and the resultant of the body's fat retains in the form of the ketone. Incapacitated metabolism of carbohydrates, malnourishment, and abnormal absorption leads to a rise in fat metabolism. To stabilize, and control diabetes mellites type 1 the control of urinary ketone comes in handy. To control diabetes mellitus type 1, urinary ketone must be tracked and maintained. Blood with a high concentration of ketones can cause dehydration, the disproportion of water-electrolyte, acidosis, and later diabetic coma. A medical condition called Ketonuria causes due to deficit of insulin in the body, can be controlled by taking insulin dosage. The presence of ketone compounds in urine variate in numbers, with 2 percent acetone, 20 percent acetoacetic acid, and 78 percent of beta-hydroxybutyric acid [11].

2.3.4 Specific Gravity

Urine specific gravity is the composition ratio of water and urine density. Urine density is directly proportional to the number of solutes. One of the major functions of the kidney is to reabsorb the water when glomerular filtration completes. It also is the most affected renal procedure by the disease. Urinalyses particularly for specific urine gravity analyze the variation of equilibrium constant anionic polyelectrolyte provided an ionized alkali medium releasing ion of hydrogen according to the magnitude of the presence of cations in the solution. The release of the number of hydrogen ions is directly proportional to the existing number of cations in the solution, which will also lower the PH level.

Variation in PH level is detected by bromothymol blue, a substance present on the testing pad. The strip only detects the presence of cations in the solution and yield a variant result (lower readings) than the result accumulated by densitometry because urine composed of a higher number of solutes that are non-ionic I.e. urea or glucose, or compounds with a higher molecular weight I.e. the materials used as radiographic contrast dispenser. Two colors, dark blue and

yellow are used to reading the results. Dark blue reads the result of 1.000 while yellow 1.030 [12].

2.3.5 Blood

Urine contains blood in either unscathed cells of red blood form called hematuria, and they are considered as a product of dismantled red blood cells known as part of hemoglobin (hemoglobinuria). Visual detection is the simplest way to detect blood provided the blood must be present in a large amount. During Hematuria, the color of urine slightly becomes reddish having blood that containing hemoglobin as a red specimen. If the test results show five or greater blood cells per microliter of urine, it needs to be treated. To find the existence of blood in urine the visual tests can't give us the absolute result, therefore, it needs to be read on a microscale I.e. microscopic testing which can detect red blood cells, but it also fails to find the hemoglobin generated by hemolytic abnormalities. Due to that chemical tests are suggested, and they are proven to be giving the most authentic results of the existence of blood. After the discovery of the blood in urine, further chemical tests are conducted for the differentiation of hematuria and hemoglobinuria. These tests can measure the amount of blood lower to the extent of five cells per microliter, but since the pad absorb the urine as well, the microscopic readings should be analyzed thoroughly. The results are categorized into three quantities. Small, moderate, and large. 1+ for small, 2+ for moderate, and 3+ for the large area used as readings of results [13].

2.3.6 pH

The variation of urine pH level ranges from 4.5 to 8 occurs due to urine production in the early part of the day carrying more acid, while the urine generated after a meal tends to be more alkaline. Specified units of measurement are used for calculating urine PH due to a wide span of variation which can provide minute details required. The body needs a balanced ratio of acid and alkali, and this function is carried out by kidneys and lungs. This ratio is balanced by excreting acidic hydrogen in many forms I.e. non-hydrogenated phosphate, weak organic acids, ammonia ions. Another way of balancing the ratio is that the bicarbonate is reabsorbed by a filtration technique called glomerular filtration that takes place in the complicated tubules of the nephron. There are two basic aims of measuring the urine PH level. The first is to diagnose and the second is to treat the disease. Not only it yields the details about the ratio of acid and alkali proportion

inpatient and spots the types of particles that are present in urine in the crystalline form but also various diseases are abolished with the decrease in PH level of urine to a recommended limit. Examples are: It removes chemotherapeutic particles, filters out specific salts as they cause the development of stones in the kidney, and limits the spread of urinary infection. The standard procedure of PH measurement has a range of 5 to 9 with an increase of 0.5 or 1 PH. Methyl red and bromothymol blue are the two chemicals used in a dual measurement testing facility. The scale of methyl red to change its color to yellow from red is PH 4 to PH 6 while it is from PH 6 to PH 9 for bromothymol blue to change to blue from yellow. The variation of colors takes place with the changes in the PH level. At PH level 5 the strip color is orange variates with the increase of PH level to yellow, green, and then dark blue at PH level 9 [14].

2.3.7 Protein

With the detection of protein, the presence of the renal disease is highly probable. A healthy range of protein's presence in urine is between 100 to 300 mg/l or excretion of protein is 100 mg per 24 hours. The glomerulus and the proteins created in the genitourinary tract filters out the proteins that are mostly comprised of low molecular weight serum proteins. The plasma majorly consists of serum protein called albumin because it has a low molecular weight. Due to the absorption of a substantial amount of filtered albumin by tubules and the presence of unfiltered albumin in the glomerulus, urine receives the lowest number of albumins. The lower number of serum and tabular micro globulins are present in other proteins. A colorimetric reaction is received as a result of the principle called protein error of indicators which is typically used by the urinalysis testing strip

Against the common notion, the Variation of the color strip not only takes place due to the change in PH level but also with the detection of protein keeping the PH level constant. Variation of colors occurs due to the acceptance of hydrogen ions by proteins. The attain more hydrogen ions the albumin accommodate a bigger number of amino groups than other proteins, that is why albumin is more reactive during the test. For different requirements, different chemicals are applied to the protein section of the strip. Bromophenol blue resides on Multisix while 3,3,5,5- tetra chlorophenol, and 3,4,5,6- tetra Bromo sulfonphthalein resides on chem strip. To control the PH level constant an acid buffer a shield of acid is used by both the chemicals. Strip color changes to yellow at PH level 3 with protein being absent. The color progression of

the strip is directly proportional to the rise in the amount of protein, initially with the lowest PH level, the strip color is orange, variates various colors with as the PH level increase reaching to green and yellow at the end. Measurement of the readings is taken in the negative; Numbers are assigned to specific quantities I.e. for 30, 100, 300, or 2000 mg/dl the numbers 1+, 2+,3+, and 4+ are assigned to each of the quantity accordingly. Analyzing trace reading is convoluted due to its values residing below 30 mg/dL [15].

2.3.8 Urobilinogen

The infused bilirubin (discharged by bile duct) is transformed into stercobilinogen and urobilinogen by bacteria present in the intestines. A little amount of urobilinogen is regathered by intestines allowing it to merge in blood and reach the liver where is it discharged. Slight amount of regathered urobilinogen, Kidneys filters out a slight amount of regathered urobilinogen that emerges in urine with a quantity of less than 1 mg/dl. Stercobilin on the other hand resides in intestines as are nonabsorbable. There is a variety of reactions that occurs on the testing strip but the two of the reactions are commonly used, I.e. Ehrlich's reaction, and 4-methoxybenzene-diazonium-tetrafluoroborate. In Ehrlich's reaction dimethylaminobenzaldehyde reacts with urobilinogen giving a variation of color ranging from light to dark pink as a result. In diazo integrating reaction, 4-methoxybenzene-diazonium-tetrafluoroborate is used giving a variation of color ranging from white to pink. This reaction is more convoluted [16].

2.3.9 Nitrates

Bacteria that decrease nitrate level in the body causes asymptomatic infections, to detect these infections a test for nitrites which is a swift screening procedure is carried out. Due to the presence of enzymes in these bacteria species, the amount of nitrate residing in urine lowers and Urinary tract infections (UTI) typically emerges due to these bacteria. Reliance on the detection of bacteria causing infections is not limited only to swift screening procedure, due to its capacity of only detecting bacteria that reduce nitrate but other procedures such as microscopic analysis or urinalysis are also utilized for diagnosis. For the detection of nitrate present in urine, a strip (pathological strip) is immersed in acid causing a reaction called Griess reaction. As a result of the reaction between nitrite and aromatic amine diazonium salt is produced, which creates pink azo dye when reacts with tetrahydroxy benzoquinone [17].

2.3.10 Leukocytes

Leukocyte esterase is found because of searching white blood cells by urinalysis. Typically, Bacteria and positive nitrite are found if the result of the nitrite test is positive. The test for leukocyte esterase cannot be relied upon for the detection of bacteria and positive nitrite as microscopic or urine culture examinations are much efficient. Leukocyte esterase Catalyze the hydrolysis of an ester of indole carboxylic acid and hence the urinalysis reaction takes place. A reaction of indoxyl (extracted as a result) reacts with diazonium salt creates azole dye of violet color [18].

2.3.11 Test Method

The initial step of the test is to submerge the testing strip in a urine sample for a calculated time, the strip after soaking is separated from the urine container making sure enough amount of urine is transferred back to urine container. Typically, a time two minutes is required for the reaction to take place and the colors start to appear on the strip depicting a particular result. The color represents a specific result according to the range provided by the lab.

2.4 Machine Learning

The storage of knowledge is converted into different models and patterns are the basis of learning. Machine learning on the other hand is collects designed data, computes the data through complex algorithms, and recognizes the patterns autonomously. Machine learning is the branch of Artificial intelligence that enables a computing machine to analyze a vast set of data, learn the patterns and predict future results, make a cluster or divide the similar future results into groups according to its pattern calculation. Majorly used Artificial intelligence concepts include Support Vector Machine, Decision Tree, Artificial Neural Network, K-means clustering, and self-organization map. Many other techniques are work in integration with the main concepts helping in yielding efficient results. Systems loaded with pre-acquired patterns are used to analyze vast data and predict results based on its intelligence of experiences.

Health-Informatics is the field of machine learning deals with the convoluted computational procedures on data of healthcare. Due to the vast databases of healthcare, it becomes humanly impossible to manipulate the data, and hence data mining is brought forward. Large chunks of data are converted into meaningful patterns of information to be used to future decision making [19].

2.5A Systematic Literature Review of Urinalysis Methods and Techniques

Urinalysis is one of the most widely used testing techniques for disease and condition diagnosis for many years. Overall, a common procedure of urinalysis is practiced most of the time in laboratories. For example, Microscopy of urine particles, chemical test strip, and urine flow cytometry [1]. Urinalysis can detect the presence of a chemical pattern in urine by using the dipstick reagents. The dipstick is immersed in the sample of urine, changing its color according to the reaction taking place in the medium, and by comparing the stick color to the chart a result is derived [2]. A urinalysis also helps in demonstrating several body functions. The diseases that Urinalysis diagnose are urinary tract infection, diseases in kidney, infection in bladder, diabetes, infectious diseases, glucosuria, and diabetes insipidus. On one hand, urinalysis can diagnose such many diseases and infections, but huge time consumption and cost are its short limitations. The other procedures are less time-consuming and less precise [3]. While many procedures and techniques are put forward in the motioned study, a research article to examine and encapsulate the latest progress in the study of urinalysis cannot be found.

This study also presents and answers the following research questions.

RQ1: What are the leading machine learning techniques currently used for urinalysis?
RQ2: What leading machine learning technologies have been used for urinalysis?
RQ3: What are the main challenges faced in urinalysis?
RQ4: Which diseases/conditions can be diagnosed by making use of urinalysis?

Several research articles are published about the procedures and techniques that can be utilized for diagnoses of diseases by Urinalysis. This study elucidates a systematic literature review (SLR) of 33 research articles published from 2007 to 2019. Alternative solutions for several problems in healthcare are suggested to be put forward and the loopholes in the current healthcare literature are addressed in this study. Resembling studies are present tackling similar problems, though no such study is conducted on statistic data of urinalysis before. Therefore, the

evaluative analysis and summarization of present research work should be conducted. The loopholes and shortcomings of current studies are thoroughly discussed, and improvements are suggested in the application of machine learning in urinalysis procedures in this study.

2.5.1 Methodology

A structured and systemized process is used for conducting a systematic literature review (SLR). A predetermined standard for the qualification has been met in the accumulation and determination of the studies. For simplification of the search six keywords are put forward, such as Strips, Urinalysis Methods, Machine Learning Application in Urinalysis, Urinalysis technologies, and Deep learning. These keywords make it easier to find specific research work.

2.5.1.1 Quality Assessment Criteria

Some predetermined standards regarding quality assurance are established to make sure that the quality of work is maintained. The research work must meet the standards.

a) Publishers

The publishers that are considered for this survey are Taylor & Francis, IEEE, ELSEVIER, SPRINGER, and ACM. Figure 1: Selected researches per publisher depicts a brief view of the papers selected from the above-mentioned publishers, also details of individual research studies for each of them is given in Table 1.

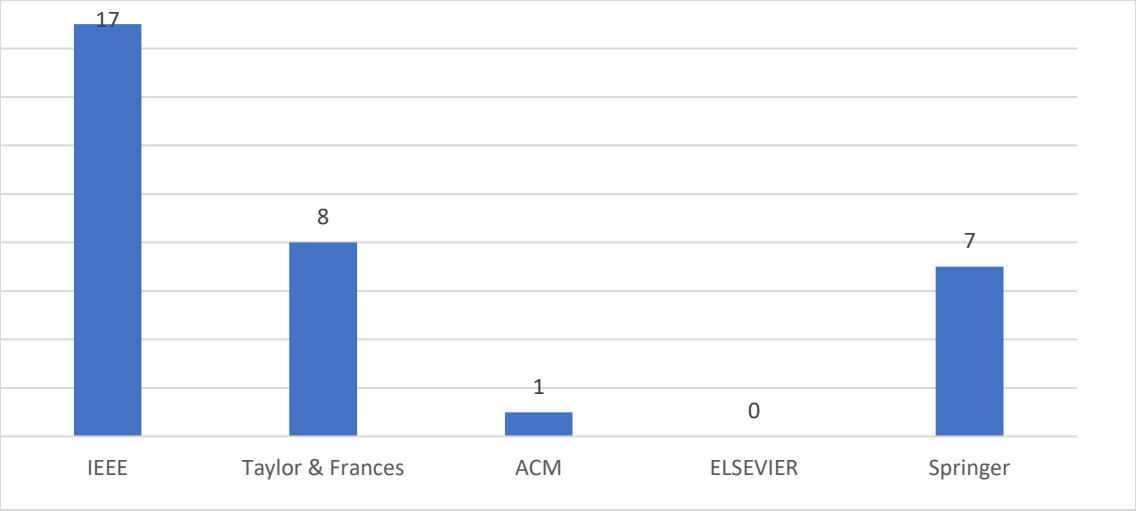


Figure 1: Selected researches per publisher

Table 1: Details of Research Studies per Publisher

Publisher	Selected Research Studies	No. of Researches
IEEE	[1] [2] [3] [4] [6] [7] [8] [9] [10] [11] [12] [13] [14] [15] [16] [17] [18]	17
Taylor & Frances	[19] [20] [21] [22] [23] [24] [25] [26]	8
ACM	[28]	1
Elsevier		0
Springer	[5] [27] [29] [30] [31] [32] [33]	7
Total		*33*

b) Presentation of an Effective Technique

The major objective of each of the papers selected is that it suggests a procedure and framework for the urinalysis. Several Papers are taken out from search results due to not fulfilling the required standards.

c) Validation of Results

The derived results from the research work should be applied to the statistical data of urinalysis to measure the validity.

d) Repetition

To be selected the research must provide, apply, and prove the authenticity of the new procedure. For reusing the old procedures, the research work is not accepted.

e) Recent Papers

To make sure that the latest papers are selected, only conference journal papers from 2007 to the present have been selected for this review as shown in Figure 2. At the end of the selection process, a total of 33 journal papers have been opted for.

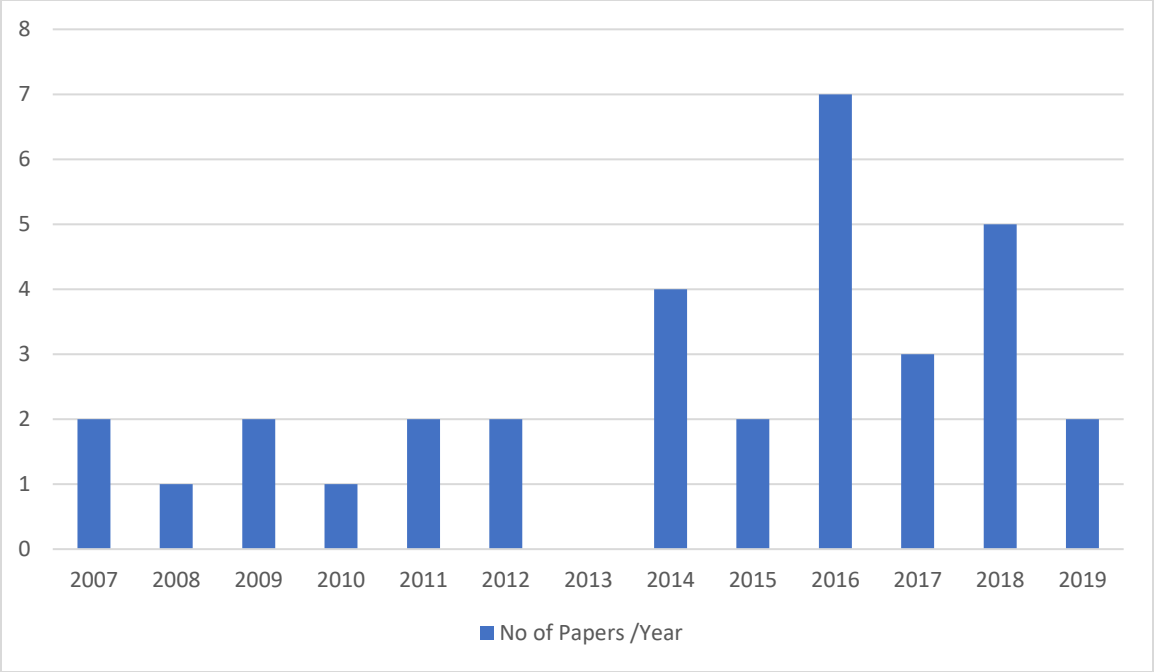


Figure 2: Selected researches per year

2.5.2 Related Work

This segment is comprised of elucidating the acquired understandings of the selected papers after a thorough examination. The main aspects of the discussion are the procedures and technologies used in these studies. The classification of the research papers has been based on procedures, diseases, hurdles, and methodologies. Several papers have put in practice the procedures to intercept the diseases and conditions such as diabetes, kidney diseases, urinary tract infections, and pregnancy in the early stage. Although several urinalysis procedures and technologies are being put forward by scientists and researchers with time, studies and research work is conducted to critically analyze the current progress in the study of urinalysis. Therefore, this study aims to critically analyze and elucidate a systematic literature review (SLR) of the urinalysis procedures that are put forward dating from 2007-2019 with a combination of 33 different studies. All these studies are comprised of 12 challenges, 10 diseases, 10 procedures, and 8 technologies.

The following are some methods that are most frequently used for urinalysis.

2.5.3 Urinalysis Methods

This section summarizes the results that are obtained through the detailed analysis of selected studies. In this article, 33studies have been identified in urinalysis. 10 urinalysis methods have been identified as given in Table 2.

Table 2: Urinalysis Methods

Methods	Papers
Adaptive Sampling Ensembles learning Method (ASELM)	[1]
Intelligent Dipstick Urinalysis Method	[2]
Flow Cytometry Method	[3]
Color Strip Test Method	[21]
Inkjet Printing Dipstick Method	[22]
Multiple Photometry Method	[23]
Stamping Method	[24]
Hand Stamped Urinalysis Method	[25]
Genetic Based Fuzzy Classifying Method	[26]
Automatic Urinary Particle Recognition Method	[27]

2.5.3.1 Adaptive Sampling Ensemble Learning Method (ASELM)

The amalgamation of ensemble learning and CSM is what adaptive sampling Ensemble learning method (ASLEM) comprised of. The analysis is carried out using various measurements which enhances the quality of statistical data. The capability of AdaBoost, Random Forest, and Bagging is analyzed and then a merged method is created to lower the chances of error in urinalysis. Ensemble learners are established to be able to conduct well than a single classifier and works vigorously in the presence of unsuitable data. The utilization of the given data will help in the construction of ensembles as well as to attain classifiers with errors that are not connected. Various methods are suggested to perform this task: (1) Utilizing the input characteristics (2) Renaming of the variant classes (3) Creating uncertainty (4) Creating subsamples of the given data to further classify and bring forward multiple solutions. Both Bagging and Adaboost's approach of operation is to intake a learning algorithm and utilize it several times on various data sets. The initial input is divided into multiple inputs, this technique called bootstrap is used in Bagging. All these inputs gather different results which are combined at the end to create a single

coherent chart. One of the good examples of such practices is the combination of Bagging and random forest. AdaBoost Algorithm follows a resampling framework in which a data of weights is managed of the initial data sets and each of the weight is modified according to the base learning algorithm. The efficiency is achieved in the AdaBoost algorithm by decreasing classification error, on the other hand, bagging decreases the variance and as a result, will be more efficient with some distortions [1].

2.5.3.2 Intelligent Dipstick Urinalysis Method

A camera is used to acquire the colors in the Intelligent Dipstick Urinalysis Method. An algorithm for modifying such as color correction, color management system, and color matching, etc. The selection of colors is based on RGB and CIELAB systems is chosen. For better color resolution and auto, the color corrector is put in practice, and a sequential linear interpolation is used to achieve a precise result of urinalysis. This process can be carried out in normal daylight by a random smartphone camera and no extra light equipment or closed acquisition box is needed [2].

2.5.3.3 Flow Cytometry Method

Inflow cytometry predefined range of color with assigned value is compared to the colors received as a result and categorized accordingly. It has the most precise results than other urinalysis procedures especially in the detection of disease if febrile patients when urine samples are utilized as the gold standard. The advantage of this technique over the other is that it is a result derivation time is less with high precision in the detection of bacteria. In this technique, the testing material is moved to a flow cell with the help of hydrodynamics, and every fragment is made sure to travel through the laser beam separately. All the information related to these particles of sample for example it's magnitude and formation is received through the calculation of the scattered beam and strength of the signal. For calculation and analyzation, all these signals are transformed into electrical signals. There are several autonomous systems of urine sample analysis in which flow cytometry can be used I.e. The Sysmex UF-1000i. Flow cytometry derives excellent outcomes with a precision that can be used to predict UTI, but the usage of variant samples o specific samples of UTI is counted as a limitation for this technique. For this study, a full-flow cytometer will be utilized such as Accuri c6 (Becton, Dickinson, and company (BD) Biosciences, Belgium).

The capability of Flow cytometry to predict positive urine cultures is tested and proven to be true by detecting and figuring out the number of bacteria with flow cytometer by our lab study recently conducted. Results concluded from the data exudes that no presence of Accuri C6 in febrile patients is found. To reduce the urine cultures, time, cost, and use of resources Flow cytometric calculations of bacteria input into practice, due to its efficient and quick finding of results of febrile patients at the ED. The major purpose of conducting such a study was to find the most precise and quick diagnosing method for urinalysis, especially in febrile patients at the ED. The efficiency and quickness in the diagnosis of disease in flow cytometry are tested and proven to be true in detecting bacteria in the urine. The customized units are flow cytometry, US, UD, and Gram stain utilized in the detection of bacteria in urine in the febrile patients at the ED, which uses a positive urine culture (105 CFU/ml) as the gold standard [3] [20].

2.5.3.4 Color Strip Test Method

Several results can be concluded with the help of a color strip test by the analysis of the color strip. It yields results with precision. The image processing technique of artificial intelligence is used to contrast the different results of the 100 samples data. No evident distinction found in the results of alpha below 0.05. Variation in the color of the strip is directly proportional to the amount of chemical present in urine. Two common methodologies are used for the analysis of strip colors:

1) The analysis of the results is carried out by humans by comparing the color results to the predefined chart. 2) The analysis of results is carried out by autonomous systems. Shortcomings of analyzing the results by the autonomous system are that it is too time-consuming as about 2 minutes are taken for calculating the result of an only single strip. To solve this problem image processing is used that will calculate many strips data at once using color recognition. The image processing technique is a good alternative to a smart autonomous system. This is preferably selected over other techniques because it gives the same results as an expert system but will low cost and in lesser time. It also is cost-effective. Comparing to the automated system's cost which is 10,000$, the image processing technique costs only 150-200. The usage of webcam in the process lowers the quality of the image due to small bandwidth. This intercession problem is solved by the technique called noise repression. Advancement is required in the system to

mobilize it easily and enhance the image quality. For mobilization to be made easy, this technology used to be made to be used in smartphones [21].

2.5.3.5 Inkjet Printing Dipstick Method

For the detecting of several diseases such as diabetes and urinary tract infections, Inkjet Dipstick can be utilized as the best option. This technique is feasible and has large implications for diagnosing major diseases such as liver diseases, diabetes, pregnancy, and preeclampsia. It is widely utilized locally in clinics and hospitals and is proven to be unerring and a good alternative to expensive tests. It is also available in mobile health units providing ease of access to the patients who need basic health care advice. Technology can provide a huge service for basic health evaluation. I.e. By the variation in solutions just printed, the substitution of the ink cartridge is carried out, which yields results for finding methanol in drinking water is being printed in the environment which is rapidly used. This ease of solution combined with the ability to change rapidly has made it stand out as a feasible technology and can move forward in the future [22].

2.5.3.6 Multiple Photometry Method

In this method, various colorimetry units are utilized to calculate the results of multi-pad dipstick for urinalysis. Several methods of photometry are used for mobilization purposes and having a simple, minimalistic design. It is a simply designed, cost-effective lab test capable of yielding good results that contain information about the urinary tract and other diseases. Its mobility and portability are what makes it able to attract the healthcare market. The latest approach for the selection of technologies for urinalysis evaluates it based on its portability, mobility, and the ability to derive results of multiple dipsticks with one photometry module. The advancement of stable photosensors and light-emitting diodes has led to the alteration in the scenario. The multi photometry system without high mobility could yield better results, be more reliable, simpler to use, and cost-effective. A multi-channel mirror module is used for the analysis of urine is designed implemented. The basic unit of this system was one colorimeter that was developed first, and later other modules added on top of it. These modules included measurement and control circuits that were connected using multiplexers for simplifying the circuits. All these modules will be utilized as a functional unit of the system therefore

connectivity between them is essential. Serial communication customization was invented because of problems that occurred due to writing directly which was consisted of a set of commands, proven, and verified by the emulation program [23].

2.5.3.7 Stamping Method

The stamping method has a far reach in the industry and invention of the healthcare industry. This system is comprised of a filter paper, a foam stamp, and it is majorly preferred for its capability of printing with high consistency, cost-effective, and easy to use. This system is used for designing a point of care of urine test strips that works like the frequently utilized test strips. The ease of use is achieved through the feature of customization of the stamping process and modularizing the number and type of units that would detect the disease, and the cost is reduced due to the cheap stamping process. To readily export these systems to other countries a minimalist, cost-effective, and mini-scale stamping procedure is invented. The cost is reduced using cheap and readily available materials. Also, recycling and reusability have lowered the costs as well. A thorough study and analysis of the properties of materials led to the development of a cost-effective, simplifying the maintainability for stamping assays onto materials. Large importers of this technology are developing countries because of its low production cost, disposal of assay through stamping, low material cost, and less complexity in the manufacturing is required. The comparative study is conducted to evaluate the material of the combination with multiple stamps to elucidate the best stamp-material combination. Production of this system compared to other wide-scale production is a less time consuming, cost-effective, and less convoluted process to deploy a manufacturing unit which is the requirement of developing countries. This will enable these countries to create screening products to diagnose diseases I.e. UTIs, diabetes, kidney diseases, and syphilis [24].

2.5.3.8 Hand Stamped Urinalysis Method

This technique involves the colorimetric determination assays for metabolism. Such as, the cheap substances that can be printed on filter paper are ketone and glucose and its comparison with other processes reveals that it has a good performance. For the assessment of diabetes, major blood testing as well as urinalysis is used. During the diagnosis process, the patients are first analyzed to know the cause of disease and then diagnosed with the disease. To diagnose diabetes

at a low cost, locally created two-parameter versions of urinalysis strips should be utilized and which is cheap in cost and efficient in performance. Hand stamped Urinalysis method uses detection reagents which are hand stamped using the foam stamps on top of the dried filter papers. These stamping papers can yield a comprehensive result of healthcare. Urinalysis test (URS-2K) tests for glucose 110, 60, 30, 15, 5 mmol/L for ketone at 8.0, 4.0, 1.5, 0.5 and 16 mmol/L. The strips react differently with different materials due to chemicals present on its surface, and a variety of color range is used for verification of image analysis. Production of these strips made it easy to the level that they can be created locally and analyzed at home, leading to the mass production in areas such as Africa. This study depicted the manufacturing methods to create strips of urinalysis that can detect diabetes makers I.e. Glucose and ketones [25].

2.5.3.9 Genetic Based Fuzzy Classifying Method

This method is used for classification. Through the help of making fuzzy logics, this method can detect and monitor the infections in the urinary tract and diseases of kidneys. The practice of generating fuzzy logic is widely used all over the world for its tremendous results in the detection process. To assess the ability of this system, various test cases have been conducted with many experiments, testing the fuzzy partitions, and proposed genetic algorithm. Other tests were conducted to check the system's capability of comprehending the logic of perturbation of tests in autonomous UF (urine flow cytometry). The Method that is used for detection is based on the principles of fuzzy genetic learning. To comprehend the mappings and patterns from the practical inputs and outputs comprised of data elucidated from applied fuzzy sets, fuzzy classification is utilized. This enables the computational system the leverage to operate fewer symbols and labels linked with the fuzzy sets. This is a widely used practice all over the world for detection and analysis. If else logic was initially provided by humans. A tremendous amount of research work is carried out to allow artificial intelligence to generate its fuzzy logic. For a rule-based fuzzy logic, a genetic algorithm is utilized to create and optimize the design. The G-A based research related to fuzzy logic systems are categorized into the repetitive process of learning, Pittsburgh, genetic collaborative determined learning, Michigan approaches. A group of data related to fuzzy logic is coded individually in the Pittsburg mechanism, while other approaches utilize fuzzy if-else rule as an individual. Every chromosome has a representation in the form of rule in the iterative rule, the

repetition of running the algorithms gathers different data and creates a unique pattern out of it, and the best solution is picked after analyzing all the gathered data. The break-in half and search rule are applied to iterative rule of learning, which lets it use the lesser time for searching for data. Slave and Mogul are the practice approaches mentioned below. MOGUL and SLAVE are proposals that follow this approach [26].

2.5.3.10 Automatic Urinary Particle Recognition Method

This method works based on the detection of objects. For searching, Convolutional Neural Network (RCNN) object detectors and single shot multibox (SSD) is used. A segmentation free task-based features are what these methods capable of. By analyzing the particles of urine using the microscopic images, the researchers can find about the presence of urinary tract and renal diseases. Human-made features for recognition are utilized by typical autonomous algorithms because manually examining the particles of urine is time-consuming and laborious. This paper is about avoiding the use of human-made features and rather using the complex neural network (CNN) which can learn the new features by itself, apply the learned features, and detect the particles present in urine. Recognizing the particles in urine is considered as object detection and put in use the commonly utilized methods such as Single shot multibox detector (SSD), Faster R-CNN as well as other particle recognition methods. Complex neural networks are used more elaborately to analyze and enhance the detection procedure. Data of about 5,376 different images comprised of 7 sets of urinary particles (leukocyte, epithelial, erythrocyte, crystal cast, myocyte, epithelial nuclei) are thoroughly analyzed and the mean average with precision (MAP) is taken which is 84% alone with the result of 72 Ms/image on the NVIDIA Titan X GPU. [27].

2.5.4 Urinalysis Technologies

This section summarizes the results that are obtained through the detailed analysis of selected studies. In this section, 8 urinalysis technologies have been identified in urinalysis as given in Table 3.

Table 3: Urinalysis Technologies

Technologies	Papers
Mobile App (Augmented Reality)	[28]
Smart Phone-Based Colorimeters	[29]

Mobile Urinalysis for Maternal Screening (MUMS)	[30]
Electronic Medical Database (ECMED)	[31]
Pocket-Sized Colorimetric Urine Reader	[32]
Strip Test App (Biochemical Test)	[20]
Point of Care testing (POCT)	[33] [34] [35] [36]
Smart bedit	[37]

2.5.4.1 Mobile App (Augmented Reality)

With the help of Augmented reality, we can easily measure the orientation and location of a test present in printed form automatically. The pattern colors that are present on the derived printed result helps calibrate the colors by itself. No hardware is required to facilitate this method, it is measure with ease and precision, and colorimetric test geometry is easily applicable to this method. The cost-effective methods for recognizing the particles include Lateral flow and fluorescent test. But colorimetric tests have an edge over these methods as it can perform a variety of diagnoses of point of care health as well as a screening of samples such as saliva, blood, and urine. The limitations of colorimetric tests are that the results are hardly put into data as the human capacity of color measurement is limited. An example of this is the conduction of multi-parameter tests which yields result in a convoluted form that is hard to comprehend. To overcome this problem the Augmented Reality is put into practice which is capable of tracking the positions and orientation of the printed test that has a pattern of colors used for creating color calibration automatically. No hardware is required to perform this solution, and the test results are easily analyzed in colorimetric test geometry as it is highly measurable. A survey of 23 Indian health workers, unaware of how to use the smartphone, was taken to get quantifiable data for analyzation. After the candidates were educated about for 10 minutes, every candidate had to manually carry out 50 different calculations. After this, the candidates had to perform the same calculations on the mobile app that we had created. Calculations performed on the mobile app were more accurate and faster with the results (R2=0.944, Time is taken = 32 sec), while calculations performed manually were less accurate and slower with the results (R2= 0.819, Time taken = 37 sec). This survey proved our mobile app to be able to calculate colorimetric tests with precision and in less time, without the dependency on hardware equipment. The

augmented reality proved to be utilized in the development of mobile tools for different medical purposes that are cost-effective and user friendly [28].

2.5.4.2 Smart Phone-Based Colorimeters

A smartphone colorimeter is a combination of a mobile smartphone app, connected to a low-cost 3D printer. Customization of the environment is a feature of this system that enables you to tailor conditions according to your requirements. E.g. the intensity of light can be changed as it is directly proportional to sensitivity. This system is categorized into two kinds of testing facilities. First is only a mobile app designed for specific calculations, and second is a combination of app and specifically designed hardware for certain functionality. These smartphone app-based colorimeters are a good alternative to the high cost and less efficient colorimetric readers. In contrast to the commercially made colorimetric readers, the only shortcoming of this system is the need of tweaking the app for it to perform the slightest task such as altering the light. The development of mobile technology has led to an improvement in the point of care diagnostic tools. In this study, an experiment of measuring the amount of protein and glucose in a biological sample was carried out, with the apparatus of a low cost, smartphone app-based colorimeter. This apparatus does not rely on external lighting therefor the tweaking to maintain balanced light is not required. This apparatus can carry out several other tests to calculate the amount of urine particle e.g. bilirubin, ketone, nitrite, urobilinogen, hemoglobin, leukocytes, and many other by programming the app and adding required calibration equations. This low-priced smartphone-based colorimeter costs a few dollars because of the presence of a smartphone in advance. This apparatus is a great alternative to commercially made colorimetric readers that cost a few hundred dollars. The extra feature of this app is that it allows you to share the results with other representatives. It is not only capable of storing the biodata of the user along with the test records but also can send this data to the cloud database that is the reason this product can be categorized as a telemedicine application. [29].

2.5.4.3 Mobile urinalysis for Maternal Screening (MUMS)

This is a mobile app-based urinalysis system designed for screening and diagnosis of maternal disorders, comprised of a lightbox. This lightbox is the major component of carrying out the urinalysis of pregnant women, costing very low. The tests help identify risk factors and warns

about issues the patient can suffer in the future in the beginning stage of pregnancy. Statistical data shows the death rate of women is quite high due to pregnancy-related complications. Proper health care facilities are not available to the poor communities especially in the developing world is the leading cause of unhealthy births of children. To tackle this problem, a low-cost screening and diagnostic tool are required that are portable and user friendly. We created a solution for fulfilling these requirements, which can be used by medical staff for finding out the complications that can occur in the future. Major components of this system are An imaging unit connected to the web application yielding the result to a gadget computer of the urinalysis test. The photo of the test strip in analyzed by the smartphone app and measures the number of biochemical substances present in the urine test sample outputs the results and sends a record to the database of the patient. The affordability of the solution has led to being accessed in poor communities, which will help them with detection, a screen of risks, and prevent pregnancy complications in the future. [30]

2.5.4.4 Electronic Medical Database (ECMED)

Biodata and medical test results data of the patient is stored in a database (ECM) which is accessible through the internet and other networks. The database is composed of different medical records of patients such as urinalysis test results data. The electronic medical database (ECMED) is accessed by patients for finding their medical records, through the internet or other networks like GSM mobile data. The electrocardiogram (ECG), and urine analysis (UA) are also stored in the patient's record. This is an autonomous urinalysis test, having the database that gets a multimedia messaging service (MMS) data of the strips, analyzes it based on the given algorithm, and outputs the urinalysis results. A short messaging service (SMS) facility is provided to the patient to report the test's results. Also, the patient's record can be updated online. The ECMED can be accessed by patients to check their records using mobile phones having access to the internet or any other networks. This automated testing facility has allowed the patients to perform the testing procedure by themselves using smartphones with GPRS. The lab admin can easily carry out diagnosing procedures by accessing the patient's data from the online database. Email service is also available that lets the patient and lab admin share information. The Telehealth system is based on the framework mentioned in research works earlier conducted but is now integrated with the advanced mobile technology. [31]

2.5.4.5 Pocket-Sized Colorimetric Urine Reader

This technology is based on the principles of detecting colorimetric. The pocket-sized colorimetric Urine Reader is comprised of a poly optical splitter (POS), a silicon photodiode (SPD), and a light-emitting diode (LED). The elaborate details of the system components are; A pocket-sized colorimetric reader of the size (2.5 cm3 x 6.5 x 13.5), a urinalysis paper strips of ten parameters e.g. (Leukocytes, glucose, erythrocytes, protein, specific gravity, glucose, pH bilirubin, urobilinogen, ketones, nitrite), a data sending capacity with the help of smartphone e.g. (Omnia 2, Samsung Electronics), and mobile using windows operating system. It has three chromatic light-emitting diodes (LED), a poly-methyl methacrylate (PMMA) optical splitter (POS, and a silicon photodiode (SPD). The technique of data reading such as transforming the signal data from red, green, blue (RGB) to hue color range or the Y model data is used, and for the measurement, the curve fitting module is implemented. Pocket-sized colorimetric is a cost-effective, portable, battery-powered, and fast processing, and analyzing reader. This reader tested in Eunji University hospital by putting into practice for analyzing thousands of urine samples, resulting in providing a precise measurement of urinary glucose and protein. Ease of access is mastered to the level that an unskilled user can operate and transfer data to the researchers and lab admins. [32]

2.5.4.6 Strip Test App (Biochemical Test)

The strip test app captures the image of strip color pads using a mobile phone cam and then autonomously analyze the images using the app that is operating in the same phone. The image taken is put against the data of pads, and if matches one of the records is categorized in pre-defined disease. This study provides the latest approaches that meet the current requirements and trends such as developing a home-based health care system which automatically takes analyze and yield result. To meet these criteria, a mobile app for strip testing that automatically analyzes the biochemical test is developed.

Following are the major advantages of this mobile App:

- Speedy analysis of the data, with automatic decision-making algorithms that lowers the error of human input to almost none.

- It is highly cost-effective, user friendly and a portable system and no external hardware equipment is required besides the strips and smartphone.
- The user has access to the results and is notified to see the clinician if the results are positive.

The advent of mobile phones and its contribution to health care has great affects e.g. (I) Mobility: This allows the patient to have access to the database every time, everywhere. (II) Reminding:

The patient is notified about the need of seeing a doctor through mails, SMS, or calls. (III) The transferring data is easy and less time-consuming on can be transferred using the internet and other networks. [20]

2.5.4.7 Point of Care testing (POCT)

This method creates an array holding different references for the urinalysis test in an innovative way. A chart by the name Doughnut chart is used to depict the categories of strip colors. Urinalysis can be carried about by human beings with the naked eye using strips without any external information. For the presentation of results, an algorithm is used which is a smartphone app.

This algorithm takes an estimation of the analyzed data from the images of strips. It can estimate numerous urine tests. With the advent of point of care testing (POCT), a fast result generating, easy to use, and portable urinalysis facility is available. This study will provide the advent of a new urinalysis test having a smartphone-based colorimetric algorithm for POCT. The aspects of advancement such as a simple design, unique idea, fast delivery, and precision in results have great implications on the performance of POCT and all the paper-based testing devices. Results of POCT depict that this system can analyze both the detection by the naked eye or smartphones. Once the process of data gathering regarding the sample, this data is transferred to the lab and research centers for further decision making. The DONUT algorithm which is independent of manual tweaking can perform in a variation of lighting environments, also in shadows. This does not only work on strip-based urinalysis but also can be utilized by other colorimetric detections e.g. drug strip tests and water quality tests. This application can be further developed

in the future and will be available on many platforms. This study provides the analysis in RGB form, which can be enhanced in CIE Lab or HSV color space. [33][34][35][36].

2.5.4.8 Smart bedit

This bedlt is a toilet stool composed of a sensor that measures urine. This device is connected to the home videophone server. A smart bedlt has a urinalysis device in it. It requires the urine strip to be put inside manually by the user. After inserting the strip, the device calculates and provides the result. This result is sent to the videophone home server via the USB/RS-232 interface. Optional measurement sensor in smart bedlt efficiently provides the parameters via the sensor network to the server. This sensor runs on NanoQplus OS [7] and works in ubiquitous sensor networks as well. Commercially built, CC 2103 Plus A with a urinalysis sensor is used in the development of this prototype by us. The home server has access to a broadband interface. Further development can be done by using broadband access to the healthcare information network, which will allow us to connect to the ambulance services, doctors, online medicine stores for smart medicine, and convoluted analysis utilizing bioinformatics schemes. [37]

2.5.5 Urinalysis Challenges

A total of twelve challenges related to Urinalysis have been recognized so far as mentioned in Table 4. These include: Presence of noise and the variance in urine data sample makes the process of analyzing, identifying, and classifying of diseases of urine complex and difficult [1]. Small data groups also lead to an increase in complexity. Many urine samples enhance the quality of results [1] Due to the use of webcam the good quality images are not available for the urine sample [20] [21] [38]. The difficulty is analyzing the data is directly proportional to the increase in the variety of samples of colorimetric which is having multiple parameters and the analysis is carried out on all simultaneously [28]. This also leads to an increase in the complexity of reading the results with the naked eye due to color and light distortion [2] [20] [29] [34] [35] [39]. Therefore, it requires manual correction which can only increase the error [20] [31] [33]. The urinalysis process is labor-intensive, time-consuming, and required microscopic examination because sometimes it becomes difficult to interpret results with the naked eye [3] [34] [40].

Table 4: Urinalysis Challenges

Challenges	Papers
Noisy and Imbalance data	[1]
Small data set	[1]
Low image quality	[20][21][38]
Test Complexity	[28]
Color Inconsistency	[2]
Light Intensity	[2][20][29][34][35][39]
User Intervention	[20][31]
Prone to human error	[33]
Difficult to interpret results with the naked eye	[34]
Need microscopic examination	[34]
Time Consuming	[39][40]
Labor intensive	[40]

2.5.6 Disease

In this segment, 10 diseases that can be diagnosed at early stages with the help of urinalysis have been identified, as given in Table 5.

Table 5: Disease Diagnosed by Urinalysis

Disease	Papers
Kidney disease	[21] [24][26][27][32][37][41][42][48]
Urinary tract infection	[21][22][23][24][25][26][27][32][39][40] [42][44][45][46]
Diabetes	[2][20][21][22][24][29][35][43][45]
Pregnancy	[20][28][29][30][39]
Syphilis	[24]
Glucosuria	[47][48][49]
Tumor	[43]
Bladder Inflammation	[43]
Infectious disease	[43]
Diabetes insipidus	[43]

2.5.7 Analysis of Urinalysis Methods and Technologies

To explore the previous research work, some studies that have been chosen, some guidelines have been set and all the selected research studies are assessed according to the guidelines. The following are the guidelines to assess the selected research studies.

2.5.7.1 Analysis Criteria for Urinalysis Methods

The analysis criteria for urinalysis models are explained as follows:

a) Has the classification been done using machine/deep learning algorithms? b) Does it take more time to perform urinalysis? c) Does it cost more? d) Does it require Labor to carry out the process? e) Does it require a microscopic examination to validate the results? f) Does it require any hardware such as i.e. urine analyzers, color detectors? g) Does it cover all of urines contents such as Bilirubin, PH level, Glucose Level, Leukocytes, Red Blood Cells, White Blood Cells, Nitrates, protein, and ketones?

Table 6: Analysis of Urinalysis Methods

Methods	**Criteria**						
	Classification	***Time Consumin***	***Cost***	***Labor-Required***	***Microscopic Examination***	***Hardware Required***	***Cover All Contents of Urine***
Adaptive Sampling Ensembles learning Method (ASELM)	Yes	Yes	Average	Yes	Yes	Yes	Yes
Color Strip Test Method	No	Yes	High	Yes	Yes	Yes	Yes
Hand Stamped Urinalysis Method	No	Yes	Average	Yes	No	Yes	No
Intelligent Dipstick Urinalysis Method	No	Yes	Average	No	Yes	Yes	Yes
Inkjet Printing Dipstick Method	No	Yes	Low	No	No	Yes	No
Multiple Photometry	No	Yes	Average	Yes	No	Yes	Yes

Method							
Stamping Method	No	Yes	Low	Yes	Yes	No	No
Flow Cytometry Method	No	Yes	High	Yes	Yes	Yes	NO
Genetic Based Fuzzy Classifying Method	Yes	No	Average	No	No	No	Yes
Automatic Urinary Partic Recognition Method	Yes	No	Average	No	No	No	Yes

2.5.7.2 Evaluation Outcome

In the given analysis of 10 methods/models, as shown in Table 6, the Automatic Urinary Particle Recognition Method is the best option because in contrast to others it allows you to perform urinalysis using machine learning classification algorithms within minimum time [27]. Besides this, it does require supportive hardware and covers all aspects of urinalysis within minimum time.

2.5.7.3 Analysis of Urinalysis Technologies

Analysis Criteria: Each suggested technology is evaluated based on; a) Machine/Deep learning-based classification b) Portability: How easily It can be carried or moved from one place to another c) Cost: How much will be the procedural cost d) Online Accessibility: Does it require online access to store or retrieve patient records in the database? e) User-Friendly GUI: How easily it can be used to perform a urinalysis and how easily it can be learned or understand f) Hardware: Does it require any additional hardware?

Table 7: Analysis of Urinalysis Technologies

Technologies	**Criteria**						
	Classification	***Portable***	***Cost***	***Online Accessibility***	***User-Friendly GUI***	***Hardware Required***	***Cover All Contents of Urine***
Mobile App (Augmented Reality)	Yes	Yes	Low	No	Yes	No	Yes
Smart Phone-Based Colorimeters	No	Yes	Average	No	Yes	No	Yes
Mobile urinalysis for Maternal Screening	No	No	High	Yes	No	Yes	No

(MUMS)							
Electronic Medical Database (ECMED)	No	No	High	Yes	No	Yes	Yes
Pocket-Sized Colorimetric Urine Reader	No	Yes	High	Yes	No	Yes	Yes
Strip Test App (Biochemical Test	No	Yes	Low	No	Yes	No	Yes
Point of Care testing (POCT)	No	Yes	Average	No	Yes	No	Yes
Smart bedit	No	No	High	No	Yes	Yes	Yes

2.5.7.4 Evaluation Outcome

The analysis of Technologies is given in Table 7. 8 technologies have been identified and analyzed that have been proposed in different studies. According to analysis, it has been found that Mobile App (Augmented Reality) has been considered the best option among all identified technologies [28]. Machine/deep learning classification techniques have been used to classify the sample and provide reliable and accurate results within time.

2.5.7.5 Future Work

When health is at stake, diseases need to be diagnosed timely to prevent them from threatening vital human life. That is the reason why these results can be improved, and hence further research is intended to be carried out to deliver better solutions to the challenges confronted by urinalysis. Besides this, a detailed analysis of known challenges will be carried out in future researches as well.

2.6 Summary

As the title of the chapter suggests, this chapter of the manuscript deals with the literature review and background study of the thesis. It starts with an illustration of pathology, which serves the purpose of assisting to determine and evaluate abnormalities in every aspect of patient health caused in the structure by the disease. Every aspect of the patient's health is taken in considering whether the procedure is based on diagnostic testing to state-of-the-art genetic technologies, to prevent disease, and to advise treatment. Then it precedes the types of pathology, which are: clinical pathology, anatomical pathology, and molecular pathology

Then it describes pathological strips, which can be used to diagnose disease by finding the changes that occurred in urine via urinalysis. The information is filtered intelligently and is presented at a suitable time to improve health or healthcare. Furthermore, it describes a standard urine test strip consists of ten chemical pads. These pads react after immersing in and removing from the urine sample. Usually, it can take from 1 to 2 minutes after dipping; some tests need a longer period though. The routine analysis of urine with a multi-parameter strip is the basic step for diagnosing a range of different diseases. Analysis can be looking for the presence of ketones, bilirubin, specific gravity, glucose, blood, protein, pH, leucocytes, and nitrate. These all are tested for infections by pathogens.

Then a brief of Machine Learning is provided, which includes artificial intelligence. Machine learning has some powerful techniques which learn huge data set and correctly classify, predict, or cluster the data set. These techniques include Decision trees, K-means clustering, Artificial Neural Networks and Support Vector Machines, etc.

After setting the tone for and providing the basic background of pathology, pathological strips, and machine learning, the chapter proceeds to present a systematic literature review of machine learning techniques applied to urinalysis. The SLR aims to find, sum up, and inspect the recent state of the art publications.

This research has been conducted systematically, to identify the latest development in urinalysis. Particularly, 33 research studies that have been published during 2007-2019 related to the specified domain have been identified and analyzed. This leads to the identification of 10 methods, 8 technologies, 12 challenges, and 10 diseases. Furthermore, an insight analysis, of the identified methods and models, reveals that the Automatic Urinary Particle Recognition Method is the most optimum option because their computational time is minimum as well as they do not require any supportive hardware. Moreover, the analysis of technologies reveals that Mobile App (Augmented Reality) is the best option among the identified technologies. The SLR has been performed to help readers acquire the work of previous researchers in the domains of urinalysis.

CHAPTER 3

EXPERIMENTATION

3.1Introduction

This chapter of the thesis discusses the initially proposed methodology which is used for experimentation. Our main data has been developed using MATLAB, which contains the images of urine test strips. A standard urine strip consists of 10 chemical pads, but in the stated dataset each urine strip image consists of 3 chemical pads or feature i.e. Glucose, PH, and Protein. Furthermore, the number of samples has been increased by adding different variance on the training & testing datasets. For validation purposes, the dataset has been split into three parts i.e. training, validation, and testing. Two optimization techniques i.e. Stochastic Gradient Descent (SGD) and ADAM for experimentation purpose are used to reduce the losses, and for learning purposes, convolution neural network is used. Besides this, an image processing technique i.e. Euclidean Distance has also been used to compare the results with the outcome of CNN. Initially, Euclidean Distance model has been evaluated on three regent strips, then segmentation has been applied on the stated dataset to breakdown the whole strip into single images. After this, our models has been tested and evaluated on segmented images.

3.2Initial Methodology

In this section, an initially proposed methodology has been illustrated, where the dataset has been developed using MATLAB, the dataset has been split into three distinct sets i.e. training, testing and validation set for assessing the performance of the algorithm, different techniques have been used for weight optimization, and a multi-layer perceptron has been used for learning and classification purpose.

3.2.1 Data

Since this research aims to diagnose urinalysis our primary dataset is urine strip images developed by MATLAB, which contains 6000 images. In the stated dataset each urine strip consists of 1 and 3 chemical pads or feature i.e. Glucose, PH, and Protein. The dataset has been

split into three parts i.e. training, validation, and testing. The number of samples has been increased by adding different noises with different variance on training and validation datasets. The table provides a list of all applied noises on the dataset.

S#	Noise	Training Data	Validation Data
1.	salt & pepper	0.001 0.003 0.004 0.005 0.006	0.007
2.	Gaussian	Mean 0.001 variance 0.001 Mean 0.003 variance 0.003 Mean 0.004 variance 0.004 Mean 0.005 variance 0.005 Mean 0.006 variance 0.006	Mean 0.001 variance 0.003
3.	Sharpen	Radius=0.01 Amount=0.01 Radius=0.02 Amount=0.02 Radius=0.03 Amount=0.03 Radius=0.04 Amount=0.04 Radius=0.05 Amount=0.05	Radius=0.02 Amount=0.01
4.	Brightness	10 20 30 40 50	45

3.2.2 Data Preprocessing

Preprocessing is a method that transforms our raw form data into clean data before feeding it to the network. It refers to the transformation which is applied to our data. The format of data should be proper to achieve better results from the applied machine learning model. Another point is that one should format the data set in a way that multiple algorithms of Machine Learning and Deep Learning are executed in a single data set, and then the best among them is selected [50].

3.2.2.1 Segmentation

Image segmentation is a commonly used technique in digital image processing and analysis to partition an image into multiple parts or regions, often based on the characteristics of the pixels

in the image. Image segmentation could involve separating foreground from background, or clustering regions of pixels based on similarities in color or shape. For example, a common application of image segmentation in medical imaging is to detect and label pixels in an image or voxels of a 3D volume that represent a tumor in a patient's brain or other organs. Several algorithms and techniques for image segmentation have been developed over the years using domain-specific knowledge to effectively solve segmentation problems in that specific application area. These applications include medical imaging, automated driving, video surveillance, and machine vision [51].

3.2.2.2 Data Representation

The data has been set into the variables i.e. x-train, x-trainable, x-test, and x-test label. The images in variable x-train are used for training our model. The X-testlabel consists of the labels of the images in the training set. A variable x-test consists of the images to test on, and the variable x-testlabel consists of the labels of the images in the test set.

X-train shape: (3600, 42, 42, 3)
X-trainlabel shape: (3600, 1)
X-test shape: (1200, 42, 42, 3)
X-testlabel shape: (1200, 1)

The shape of the x-train data set is a 4-Dimensional array with 3600 rows of 42 x 42-pixel image with depth = 3 (RGB) where R is Red, G is Green, and B is Blue. The x-trainable data shape is a 2-Dimensional array with 3600 rows and 1 column. The shape of the x-test data set is a 4-Dimensional array with 1200 rows of 3 x 42 x 42-pixel image with depth = 3 (RGB). The x-testlabel data shape is a 2-Dimensional array with 1200 rows and 1 column. Before data normalization, the dataset has been transforming into a 2-D array. Because the convolution neural network takes a 2-D array as an input [52].

3.2.2.3 Data Normalization

The dataset consisting of 3600 RGB images has been used for the training of our model. These images were divided into blocks of 42 x 42 and then the neural network was trained on these blocks of images. These blocks statistically co-related and hence can be extended to many

images. The pixel value of the images falls in the range between 0 and 255. To validate the proposed technique, the dataset has been sectioned into two i.e. training and testing, 80% of the data has been used for the former purpose, and 20% for the latter. A normalizing function *f1* is applied to normalize the pixels values of the images between the range of 0 and 1. The pixel values for an RGB image are usually between 0 and 255. But it is usual to scale the input values of the network to a certain range. In this case, the input value should be scaled to a value of type float 32 within an interval of 0 and 1[53].

3.2.2.4 Data Reshaping

Convolutional neural networks are mostly designed in such a way that only photos of a fixed size can be input to it. This results in some problems during the collection of data and the deployment of the model. Normally to overcome this problem, the input images are reshaped to be fed in the networks [54].

3.2.3 Validation

Validation is a very important part of the process to check the feasibility and accuracy of a technique. We use the validation process right after the data preprocessing takes place [60]. We have opted for a Simple K-Fold cross-validation technique that split our data into K parts for training, validation, and testing. To train and test our model, our data should be divided into three different datasets, consists of a training set, a validation set, and a test set. First, the training set is used to train the model. During each epoch, our model will get trained iteratively on the very same data in our training and the model will continue learning about this data. Later, we can establish our model, and it will accurately predict new data labels that are never seen by this model before, and those predictions are based on the model train with current data.

The validation set in not part of the training data set; it is used to validate our model during training. The validation process helps us in adjusting our hyper-parameters for correct predictions. So, it is the same as the previously described process, with each epoch during the model training; the model will be trained on training data set and simultaneously validating on the validation data set. During the process of training the model, the model will differentiate the output for every input in the training data set. After this process of classification, the loss will be calculated and the weights assigned in the model will be modified and then in the next epoch, it

will classify this input once more. This time as mentioned earlier, again training the model will be arranging each input from the validation data set too. This classification will be based on what has been learned regarding the trained data set. The weights will be adjusted in the model based on the loss computed by our validation data. One of the major purposes of a validation set is to make sure that the model is not overfitting to the training data set.

Test data is unlike the training and the validation set. When our model training and validation using the training and validation sets are done, the model will be used to predict the outcome of the test data set. The main difference among the test set and the two other sets (training and validation), should not be given labels and the must label data sets will be training and validation to check some metrics like accuracy and loss from each epoch [55].

3.2.5 Convolutional Neural Network

The type of deep neural network used for image compression in our proposed framework is Convolutional Neural Network (CNN). A **Convolutional Neural Network (ConvNet/CNN)** is a Deep Learning algorithm that can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image, and be able to differentiate one from the other. The pre-processing required in a ConvNet is much lower as compared to other classification algorithms. While in primitive methods filters are hand-engineered, with enough training, ConvNets can learn these filters/characteristics. The architecture of a ConvNet is analogous to that of the connectivity pattern of Neurons in the Human Brain and was inspired by the organization of the Visual Cortex. Individual neurons respond to stimuli only in a restricted region of the visual field known as the Receptive Field. A collection of such fields overlaps to cover the entire visual area.

Here also the CNNs developed, consisting of three layers, working as auto-encoder neural networks. Moreover, the output layer neurons are adjusted accordingly to get higher and higher compression rates. The dataset consisting of 3600 images has been used for the training of our models. These images were divided into blocks of 42 x 42 and then the convolutional neural network was trained on these blocks of images. These blocks are statistically co-related and hence can be extended to many images. The pixel value of the images falls in the range between 0 and 255. To validate the proposed technique, the dataset has been divided into two parts i.e.

training and testing, 80% of the data has been used for the former purpose, and 20% for the latter. A normalizing function is applied to normalize the pixels values of the images between the range of 0 and 1. The pixel values for a greyscale image are usually between 0 and 255 [56].

3.2.5.1 Activation Function

To know about the working of the Artificial Neural Networks, we must know about the functionality of an artificial neuron. Neurons calculate the weighted sum of their inputs adds a bias to it and decide whether they should be fired or not. The value of Y can range from +inf to –inf, and the neuron cannot bound the value. Then how is the decision of whether to fire or not made? This is where activation functions come into play. There are numerous kinds of activation functions, namely: linear function, a sigmoid function, *SoftMax* function, and ReLU, etc. However, the selection of these activation functions varies from case to case. In our case, we have used two different activation functions i.e. ReLU and SoftMax [57].

3.2.5.1.1 Rectified Linear Units or ReLU

The main function of ReLU is to ensure the output value no to be negative which means that if *z* is greater than zero the output will remain *z*, and when the output is below zero the output remains zero. Moreover, it is also used when there are numerous output possibilities. In our case, ReLU is used for going from the input layer to the hidden layer [58].

What gives ReLU an edge is that it does not activate the entire set of neurons at once. E.g. when a negative input is received it will be converted into zero, and the neuron will not get activated. In other words, at a given point in time, a few neurons will be active and hence ReLU helps in making the artificial neural network sparse and hence increases its efficiency [59].

3.2.5.1.2 SoftMax

Without vanishing the training process with apply ReLU, the sigmoid function can be easily applied. But the classification problem cannot be tackled much. Simply speaking, the sigmoid function can only handle two classes, which is not what we expect.

The SoftMax function squashes the outputs of each unit to be between 0 and 1, just like a sigmoid function. But it also divides each output such that the total sum of the outputs is equal to

1 (check it on the figure above). The output of the SoftMax function is equivalent to a categorical probability distribution, it tells you the probability that any of the classes are true.

Mathematically the SoftMax function is shown below, where z is a vector of the inputs to the output layer (if you have 10 output units, then there are 10 elements in z). And again, j indexes the output units, so j = 1, 2, ..., K [60].

3.2.5.2 Optimization

For weight optimization normally the forward and backward propagation techniques are used. However, in our case, we use the Adaptive moment estimation technique. However, we use the following technique for optimization.

3.2.5.2.1 **Adam**

Adam (Adaptive Moment Estimation) can be utilized, available for deep neural networks. Adam is one of the best methods that are used to calculate adaptive learning rates for every parameter. It computes the adaptive learning rates for all the parameters. Apart from storing the exponentially decaying averages of previously squared gradients, for instance, RMSporp and Adadelta, it also keeps something like momentum. If momentum is thought of like a ball going down a slope, Adam can be termed as a heavy ball having friction and hence providing us with flat minima [61].

3.2.5.3 Loss Function

A group of objective functions minimized is known as “loss functions” in machine learning. A loss function determines how good the results of the prediction models are and how close the predicted output and actual output is. The target for the neural networks was to produce coefficients with little error.

3.2.5.3.1 Mean Squared Error

The most used function for regression is the Mean squared error (MSE). The loss is the mean overseen data of the squared differences between true and predicted values or writing it as a formula.

MSE is sensitive towards outliers and given several examples with the same input feature values, the optimal prediction will be their mean target value. This should be compared with Mean Absolute Error, where the optimal prediction is the median. MSE is thus good to use if you believe that your target data, conditioned on the input, is normally distributed around a mean value, and when it is important to penalize outliers extra much. Use MSE when doing regression, believing that your target, conditioned on the input, is normally distributed, and want large errors to be significantly (quadratically) more penalized than small ones [62].

3.3 Euclidean Distance

Euclidean Distance is the final technique in our experimentation for diagnosing urinalysis problem. CIELAB Euclidean Distance function has been used to calculate the difference. The difference or distance between two colors is a metric of interest in color science. It allows a quantified examination of a notion that formerly could only be described with adjectives. Quantification of these properties is of great importance to those whose work is color critical. Common definitions make use of the Euclidean distance in a device-independent color. In mathematics, the Euclidean distance or Euclidean metric is the "ordinary" straight line distance between two points in Euclidean space. With this distance, Euclidean space becomes a metric space. The associated norm is called the Euclidean norm [63].

$$d(\mathbf{p}, \mathbf{q}) = d(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

Equation 1: Euclidean Distance function

The position of a point in a Euclidean *n*-space is a Euclidean vector. So, p and q may be represented as Euclidean vectors, starting from the origin of the space (initial point) with their tips (terminal points) ending at the two points. The Euclidean norm, or Euclidean length, or magnitude of a vector measures the length of the vector.

3.3.1 CIELAB

In CIELAB color space we can visualize the difference between the source and destination gamut quantitatively. Unfortunately, the Euclidean distance in RGB space does not match the perceived color distances, and thus RGB is not well suited for gamut mapping. Therefore, two-color spaces, CIELUV, and CIELAB have been recommended by the CIE (1978). Approximate

correlates with the perceived lightness, chroma, and hue of a stimulus that can be easily derived from their coordinates. Although originally both spaces were recommended, CIELAB is almost universally used today, for color measurements [64]. In CIELAB the psychometric lightness *L* is defined as a difference in light, *a* is (red-green) and *b* is (yellow-blue) coordinates. the CIELAB color difference formula is defined as:

$$\Delta E^*_{ab} = \sqrt{(L^*_2 - L^*_1)^2 + (a^*_2 - a^*_1)^2 + (b^*_2 - b^*_1)^2}.$$

Equation 2: CIELAB

3.4Model Accuracy

Two models i.e. Euclidean Distance CIELAB and regression model are used to tackle our problem. The regression model uses Convolutional Neural Network for classification. MAE metrics are used for measuring the accuracy and loss of the models i.e. the lower the mean absolute error between the targeted and actual results the higher the accuracy and the lower the loss. The accuracy vs noise variance graphs for the models has been presented in **Figure 5-10** in chapter 5. It is clear from the graphs, that the regression model with CNN classifier has a higher accuracy as compared with Euclidean Distance results.

3.5Summary

In this chapter, experimentations have been performed to assess the methodology proposed in the previous chapter. Besides this, the results of our proposed methodology have also been evaluated and compared with the previously used method like Euclidean Distance The main dataset used in this manuscript contains the urine test strip RGB images, which is developed using MATLAB. In the stated dataset each urine strip image consists of 1 chemical pad or feature i.e. Glucose, PH, and Protein. Different noises i.e. Gaussian and salt & pepper have been applied on the dataset images to increase numbers. Initially, the dataset was comprised of three regent color strip images, which is used as an input for Euclidean Distance method. After this, segmentation has been applied to the images and then the segmented images have been used for experimentation purposes in both methods. The accuracy vs noise variance graphs for the models has been presented in Figure 5-**10**. The experimentation results reveal that the regression model with the CNN classifier has a higher accuracy as compared with Euclidean Distance.

CHAPTER 4

PROPOSED METHODOLOGY AND IMPLEMENTATION

4.1Introduction

In this chapter, the final methodology has been discussed. It is very clear from the results that the Convolutional Neural Network is the best one and hence CNN has been chosen in the final methodology. For validation purposes, 80% of the data has been used for training and the remaining 20% data has been used for testing.

4.2Proposed Methodology

A flow chart that depicts the proposed methodology has been presented in the following figure i.e. Figure 3.

4.2.1 Data

Since this research aims to diagnose urinalysis our primary dataset is urine strip images developed by MATLAB, which contains RGB images. In the stated dataset each urine strip consists of 1 or 3 chemical pads or feature i.e. Glucose, PH, and Protein.

4.2.2 Data Preprocessing

Preprocessing is a method that transforms our raw form data into clean data before feeding it to the network. It refers to the transformation which is applied to our data. The format of data should be proper to achieve better results from the applied machine learning model. Another point is that one should format the data set in a way that multiple algorithms of Machine Learning and Deep Learning are executed in a single data set, and then the best among them is selected [50].

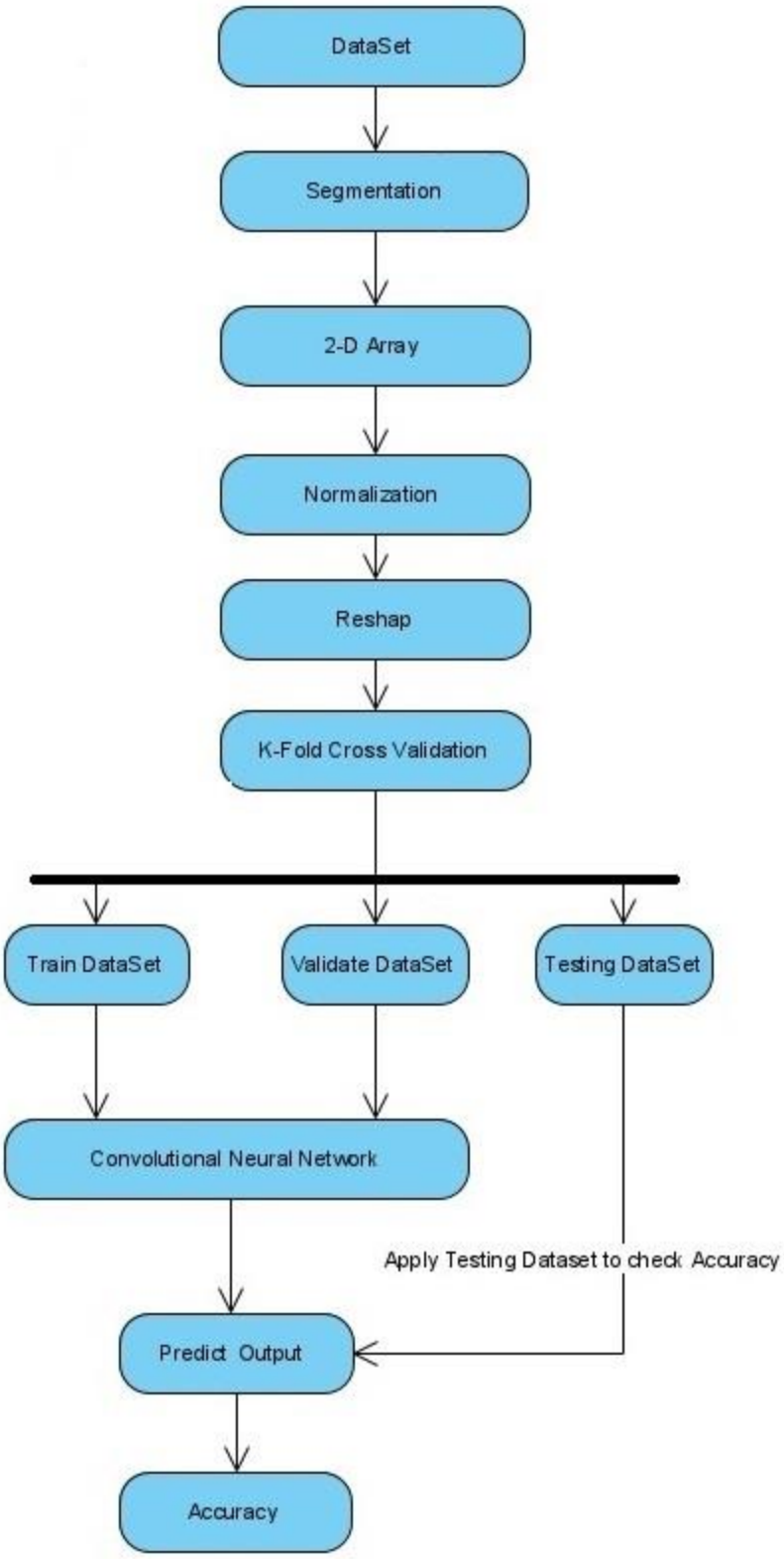


Figure 3: Proposed Methodology

4.2.2.1 Segmentation

Image segmentation is a commonly used technique in digital image processing and analysis to partition an image into multiple parts or regions, often based on the characteristics of the pixels in the image. Image segmentation could involve separating foreground from background, or clustering regions of pixels based on similarities in color or shape. For example, a common application of image segmentation in medical imaging is to detect and label pixels in an image or voxels of a 3D volume that represent a tumor in a patient's brain or other organs. Several algorithms and techniques for image segmentation have been developed over the years using domain-specific knowledge to effectively solve segmentation problems in that specific application area. These applications include medical imaging, automated driving, video surveillance, and machine vision [51].

4.2.2.2 Data Representation

The data has been set into the variables i.e. x-train, x-trainable, x-test, and x-test label. The images in variable x-train are used for training our model. The X-testlabel consists of the labels of the images in the training set. A variable x-test consists of the images to test on, and the variable x-testlabel consists of the labels of the images in the test set.

X-train shape: (3600, 42, 42, 3)
X-trainlabel shape: (3600, 1)
X-test shape: (1200, 42, 42, 3)
X-testlabel shape: (1200, 1)

The shape of the x-train data set is a 4-Dimensional array with 3600 rows of 42 x 42-pixel image with depth = 3 (RGB) where R is Red, G is Green, and B is Blue. The x-trainable data shape is a 2-Dimensional array with 3600 rows and 1 column. The shape of the x-test data set is a 4-Dimensional array with 1200 rows of 3 x 42 x 42-pixel image with depth = 3 (RGB). The x-testlabel data shape is a 2-Dimensional array with 1200 rows and 1 column. Before data normalization, the dataset has been transforming into a 2-D array. Because the convolution neural network takes a 2-D array as an input [52].

4.2.2.3 Data Normalization

The dataset consisting of 3,600 RGB images has been used for the training of our model. These images were divided into blocks of 42 x 42 and then the neural network was trained on these blocks of images. These blocks statistically co-related and hence can be extended to many images. The pixel value of the images falls in the range between 0 and 255. To validate the proposed technique, the dataset has been sectioned into two i.e. training and testing, 80% of the data has been used for the former purpose, and 20% for the latter. A normalizing function *f1* is applied to normalize the pixels values of the images between the range of 0 and 1. The pixel values for an RGB image are usually between 0 and 255. But it is usual to scale the input values of the network to a certain range. In this case, the input value should be scaled to a value of type float 32 within an interval of 0 and 1[53].

4.2.2.4 Data Reshaping

Convolutional neural networks are mostly designed in such a way that only photos of a fixed size can be input to it. This results in some problems during the collection of data and the deployment of the model. Normally to overcome this problem, the input images are reshaped to be fed in the networks [54].

4.2.3 Validation

Validation is a very important part of the process to check the feasibility and accuracy of a technique. We use the validation process right after the data preprocessing takes place [60]. We have opted for a Simple K-Fold cross-validation technique that split our data into K parts for training, validation, and testing. To train and test our model, our data should be divided into three different datasets, consists of a training set, a validation set, and a test set. First, the training set is used to train the model. During each epoch, our model will get trained iteratively on the very same data in our training and the model will continue learning about this data. Later, we can establish our model, and it will accurately predict new data labels that are never seen by this model before, and those predictions are based on the model train with current data.

The validation set in not part of the training data set; it is used to validate our model during training. The validation process helps us in adjusting our hyper-parameters for correct predictions. So, it is the same as the previously described process, with each epoch during the

model training; the model will be trained on training data set and simultaneously validating on the validation data set. During the process of training the model, the model will differentiate the output for every input in the training data set. After this process of classification, the loss will be calculated and the weights assigned in the model will be modified and then in the next epoch, it will classify this input once more. This time as mentioned earlier, again training the model will be arranging each input from the validation data set too. This classification will be based on what has been learned regarding the trained data set. The weights will be adjusted in the model based on the loss computed by our validation data. One of the major purposes of a validation set is to make sure that the model is not overfitting to the training data set.

Test data is unlike the training and the validation set. When our model has been trained and validated using the training and validation sets, then the model will be used to predict the outcome of the test data set. The main difference among the test set and the two other sets (training and validation), should not be given labels and the must label data sets will be training and validation to check some metrics like accuracy and loss from each epoch [55].

4.2.4 Convolutional Neural Network

The type of deep neural network used for image compression in our proposed framework is Convolutional Neural Network (CNN). A **Convolutional Neural Network (ConvNet/CNN)** is a Deep Learning algorithm that can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image, and be able to differentiate one from the other. The pre-processing required in a ConvNet is much lower as compared to other classification algorithms. While in primitive methods filters are hand-engineered, with enough training, ConvNets can learn these filters/characteristics. The architecture of a ConvNet is analogous to that of the connectivity pattern of Neurons in the Human Brain and is inspired by the organization of the Visual Cortex. Individual neurons respond to stimuli only in a restricted region of the visual field known as the Receptive Field. A collection of such fields overlaps to cover the entire visual area.

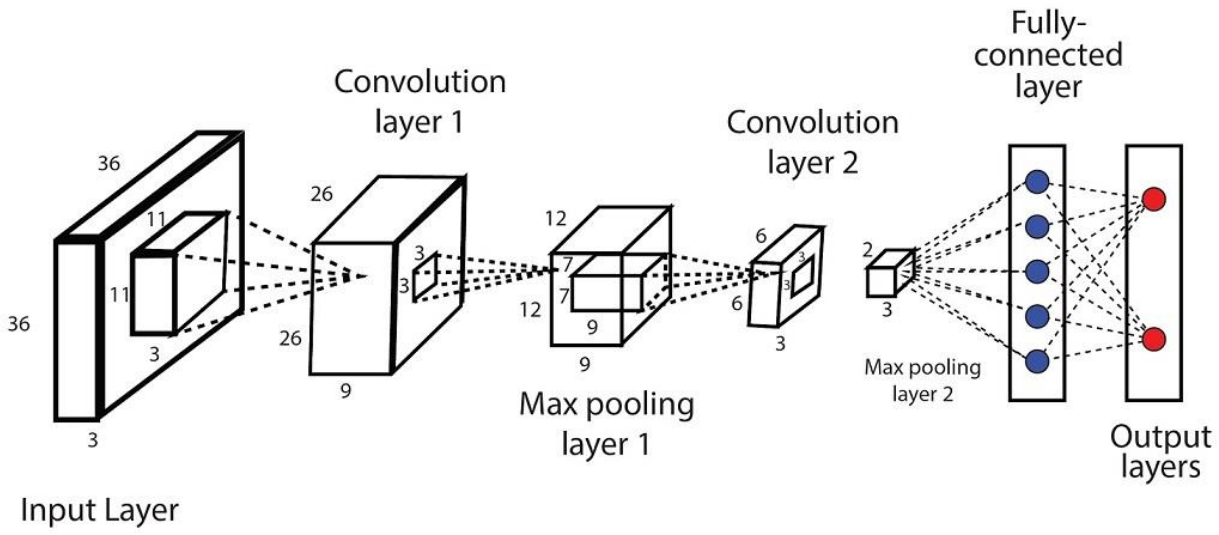


Figure 4: CNN Architecture

Here also the CNNs developed, consisting of three layers, working as auto-encoder neural networks. Moreover, the output layer neurons are adjusted accordingly to get higher and higher compression rates. The deep Convolutional Neural Networks' architecture, developed for image compression, is shown in Figure 4. The dataset consisting of 3600 images has been used for the training of our models. These images were divided into blocks of 42 x 42 and then the convolutional neural network was trained on these blocks of images. These blocks are statistically co-related and hence can be extended to many images. The pixel value of the images falls in the range between 0 and 255. To validate the proposed technique, the dataset has been divided into two parts i.e. training and testing, 80% of the data has been used for the former purpose, and 20% for the latter. A normalizing function *f1* is applied to normalize the values of the pixels of the images between the1 and 0. The values of pixels for the greyscale image are usually between 0 and 255 [56].

4.2.4.1 Activation Function

To understand what activation functions, do, we need to understand the functionality of an artificial neuron. Neurons compute the sum of their inputs data adds a bias to it to find whether which should be discharged or not. The value of Z can range from +inf to – inf, and the neuron cannot bound the value. Then how is the decision of whether to fire or not made? This is where activation functions come into play. There are numerous kinds of activation functions, namely: linear function, a sigmoid function, *SoftMax* function, and ReLU, etc. However, the selection of

these activation functions varies from case to case. In our case, we have used two different activation functions i.e. ReLU and SoftMax [57].

4.2.4.1.1 Rectified Linear Units or ReLU

ReLU makes sure that the output does not become a negative value. So, the output says that when z is maximum as compared to zero, and if it goes below zero the output stays zero. Moreover, it is also used when there are numerous output possibilities. In our case, ReLU is used for going from the input to the hidden layer [58].

$$(z) = \max(0, z)$$

Equation 3: Rectified Linear Units or ReLU

What gives ReLU an edge is that it does not activate the entire set of neurons at once. E.g. when a negative input is received it will be converted into zero, and the neuron will not get activated. In other words, at a given point in time, a few neurons will be active and hence ReLU helps in making the artificial neural network sparse and hence increases its efficiency [59].

4.2.4.1.2 SoftMax

The sigmoid function can be applied easily, the ReLU will not vanish the effect during your training process, but if you want to tackle the problems of classification, it can't be done with this which means only two classes can be handled by the sigmoid function, that's not what we want. The outputs of each unit are squashes out among 0 and 1 by SoftMax function, like sigmoid, but the extra thing is that each output is divided in a way that the sum of all outputs remains equal to 1. The SoftMax function output is like a categorical probability distribution, which describes the probability of classes with true value. The SoftMax function mathematical form is described below, z is an inputs vector for layers of output i.e. if outputs units are 5, elements in z are 5, and j indexes the output units i.e. j = 1, 2, 3, .., K [60].

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$$

Equation 4: SoftMax

4.2.4.2 Optimization

Usually, forward propagation and backward propagation techniques are used as error functions or weight optimization. But in our case, we use the Adaptive moment estimation technique. However, we use the following technique for optimization.

4.2.4.2.1 Adam

Since 2014, a special optimization algorithm in the shape of Adam (Adaptive Moment Estimation) for deep neural networks is present. Adam is one of the best methods that are used to calculate adaptive learning rates for every parameter. It computes the adaptive learning rates for all the parameters. Apart from storing the exponentially decaying averages of previously squared gradients, for instance, RMSporp and Adadelta, it also keeps something like momentum. If momentum is thought of like a ball going down a slope, Adam can be termed as a heavy ball having friction and hence providing us with flat minima [61].

4.2.4.3 Loss Function

In machine learning, 'loss functions' are a group of objective functions that are minimized. A loss function determines how good the results of the prediction models are and how close the predicted output and actual output is. The target for the neural networks was to produce coefficients with little error.

4.2.4.3.1 Mean Squared Error

Mean squared error (MSE) is the most used loss function for regression. The loss is the mean overseen data of the squared differences between true and predicted values or writing it as a formula.

$$L(y, \hat{y}) = \frac{1}{N} \sum_{i=0}^{N} (y - \hat{y}_i)^2$$

Equation 5: Mean squared error

MSE is sensitive towards outliers and given several examples with the same input feature values, the optimal prediction will be their mean target value. This should be compared with Mean Absolute Error, where the optimal prediction is the median. MSE is thus good to use if you believe that your target data, conditioned on the input, is normally distributed around a mean value, and when it is important to penalize outliers extra much. Use MSE when doing regression, believing that your target, conditioned on the input, is normally distributed, and want large errors to be significantly (quadratically) more penalized than small ones [62].

4.3Summary

This chapter of the manuscript presents the main idea and the main methodology proposed in the thesis. It builds on the experimentation that has been performed in the previous chapter i.e. in the third chapter, experimentations have been performed to assess the methodology proposed in the literature. Besides this, the results of our proposed methodology have also been evaluated and compared with the previously used method like Euclidean Distance. The main dataset used in this manuscript contains the urine test strip RGB images, which is developed using MATLAB. Different noises i.e. Gaussian and salt & pepper have been applied on the dataset images to increase numbers. Initially, the dataset was comprised of three regent color strip images, which is used as an input for both methods.

CHAPTER 5

RESULTS, COMPARISON, AND ANALYSIS

5.1Introduction

In this chapter, the results of the proposed methodology have been analyzed and discussed. The accuracy vs noise variance graphs for the models has been presented in Figure 5-**10**. The experimentation results reveal that the regression model with CNN classifier has a higher accuracy as compared with Euclidean Distance CIELAB.

5.2Model Accuracy

Two models i.e. Euclidean Distance CIELAB and a regression model are used to tackle our problem. The regression model uses a Convolutional Neural Network for classification. MSE metrics are used for measuring the accuracy and loss of the models i.e. the lower the mean error between the targeted and actual results the higher the accuracy and the lower the loss. The accuracy vs noise variance graphs for the models has been presented in Figure 5-**10**. It is clear from the graphs, that the regression model with CNN classifier has a higher accuracy as compared with Euclidean Distance results.

The following figure i.e. **Figure** 5 is used to represent accuracy vs noise graph. In this figure, the line with a square marker or Single-EC represents the accuracy for glucose segmented images with Gaussian noise using the Euclidean Distance method. The line with triangle marker or 3Strip-EC represents accuracy for whole strip images with Gaussian noise using the Euclidean Distance method. The line with a diamond marker or Single-CNN represents accuracy for glucose segmented images with Gaussian noise using the CNN classifier. It is clear from the graph that accuracy has decreased as the variance of noise increases. Single CNN shows high accuracy as compared to the single-EC and 3strip-EC.

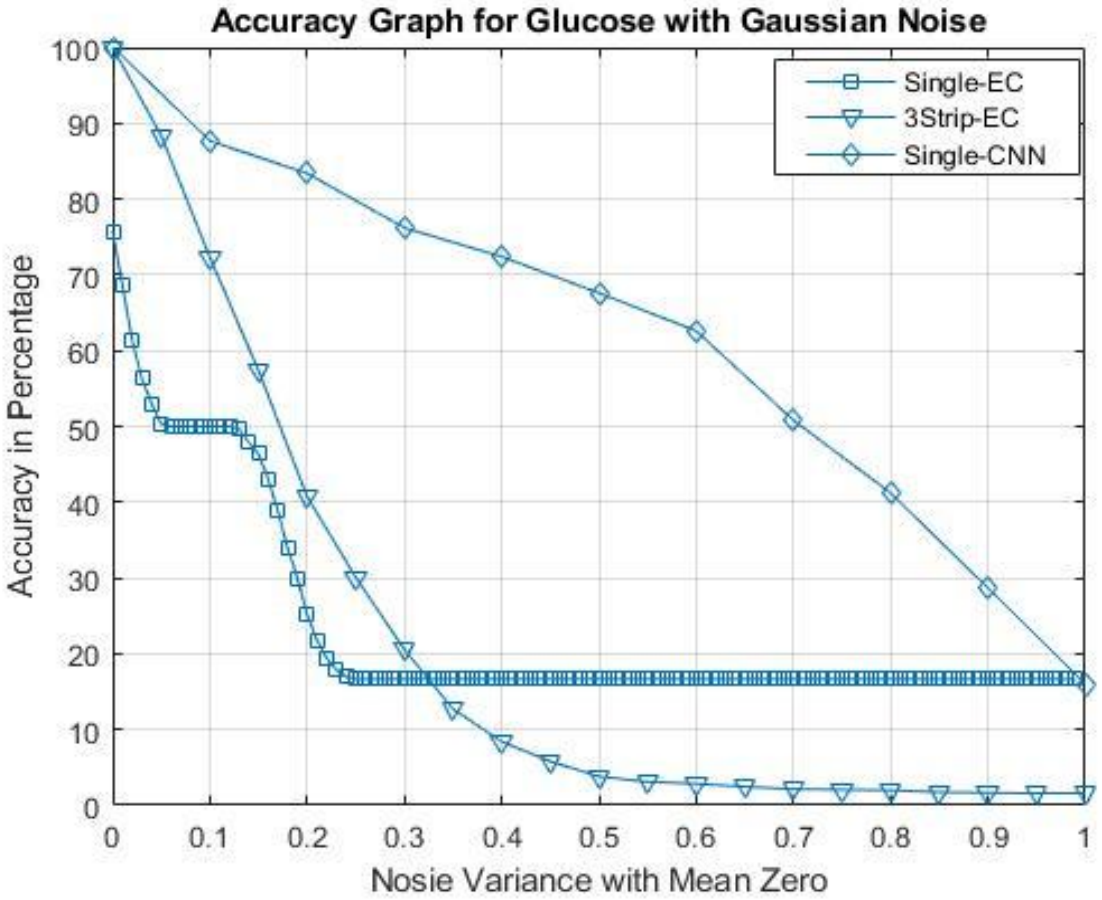


Figure 5: Accuracy Graph for Glucose & 3Strip with Gaussian Noise

The following figure i.e. Figure 6 is used to represent accuracy vs noise graph. In this figure, the line with a square marker or Single-EC represents the accuracy for glucose segmented images with Salt & Pepper noise using the Euclidean Distance method. The line with triangle marker or 3Strip-EC represents accuracy for whole strip images with Salt &Pepper noise using the Euclidean Distance method. The line with a diamond marker or Single-CNN represents accuracy for glucose segmented images with Salt & Pepper noise using CNN classifier. It is clear from the graph that accuracy has decreased as the variance of noise increases. Single CNN shows high accuracy as compared to the single-EC and 3strip-EC.

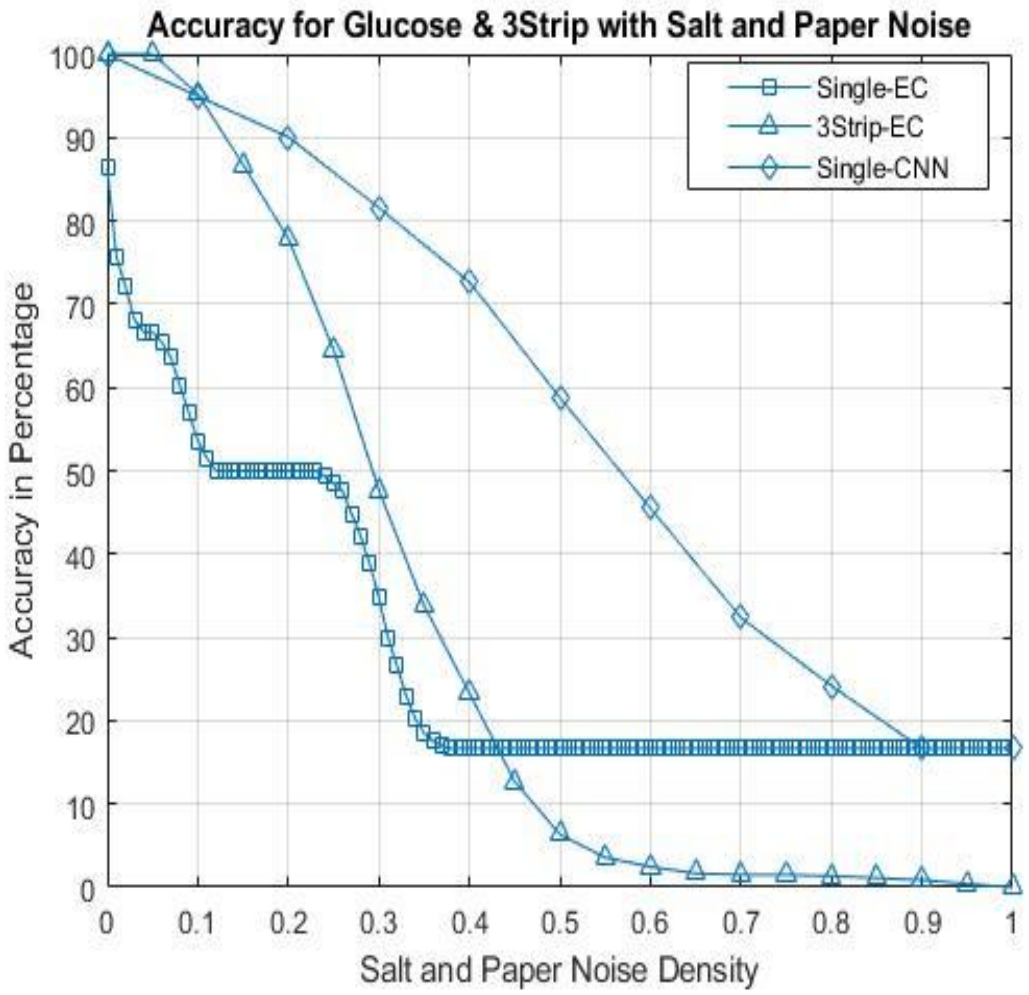


Figure 6: Accuracy Graph for Glucose & 3Strip with Salt & Pepper Noise

The following figure i.e. Figure 7 is used to represent accuracy vs noise graph. In this figure, the line with a square marker or Single-EC represents the accuracy for PH segmented images with Gaussian noise using the Euclidean Distance method. The line with triangle marker or 3Strip-EC represents accuracy for whole strip images with Gaussian noise using the Euclidean Distance method. The line with a diamond marker or Single-CNN represents accuracy for PH segmented images with Gaussian noise using CNN classifier. It is clear from the graph that accuracy has decreased as the variance of noise increases. Single CNN shows high accuracy as compared to the single-EC and 3strip-EC.

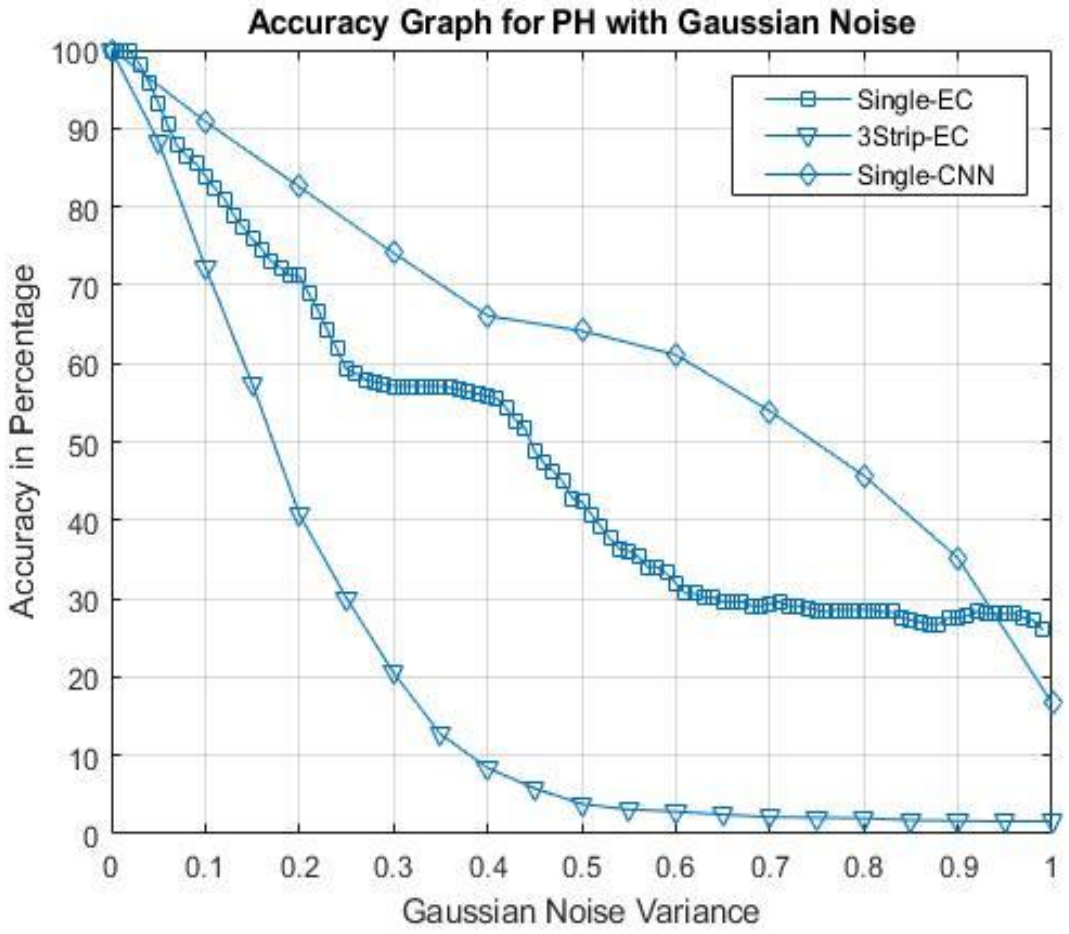


Figure 7: Accuracy Graph for PH & 3Strip with Gaussian Noise

The following figure i.e. Figure 8 is used to represent accuracy vs noise graph. In this figure, the line with a square marker or Single-EC represents the accuracy for PH segmented images with Salt & Pepper noise using the Euclidean Distance method. The line with triangle marker or 3Strip-EC represents accuracy for whole strip images with Salt &Pepper noise using the Euclidean Distance method. The line with a diamond marker or Single-CNN represents accuracy for PH segmented images with Salt & Pepper noise using CNN classifier. It is clear from the graph that accuracy has decreased as the variance of noise increases. In this graph initially at low-density noise the results of Euclidean distance comparatively better than CNN classifier but as noise increases Single CNN shows high accuracy as compared to the single-EC and 3strip-EC.

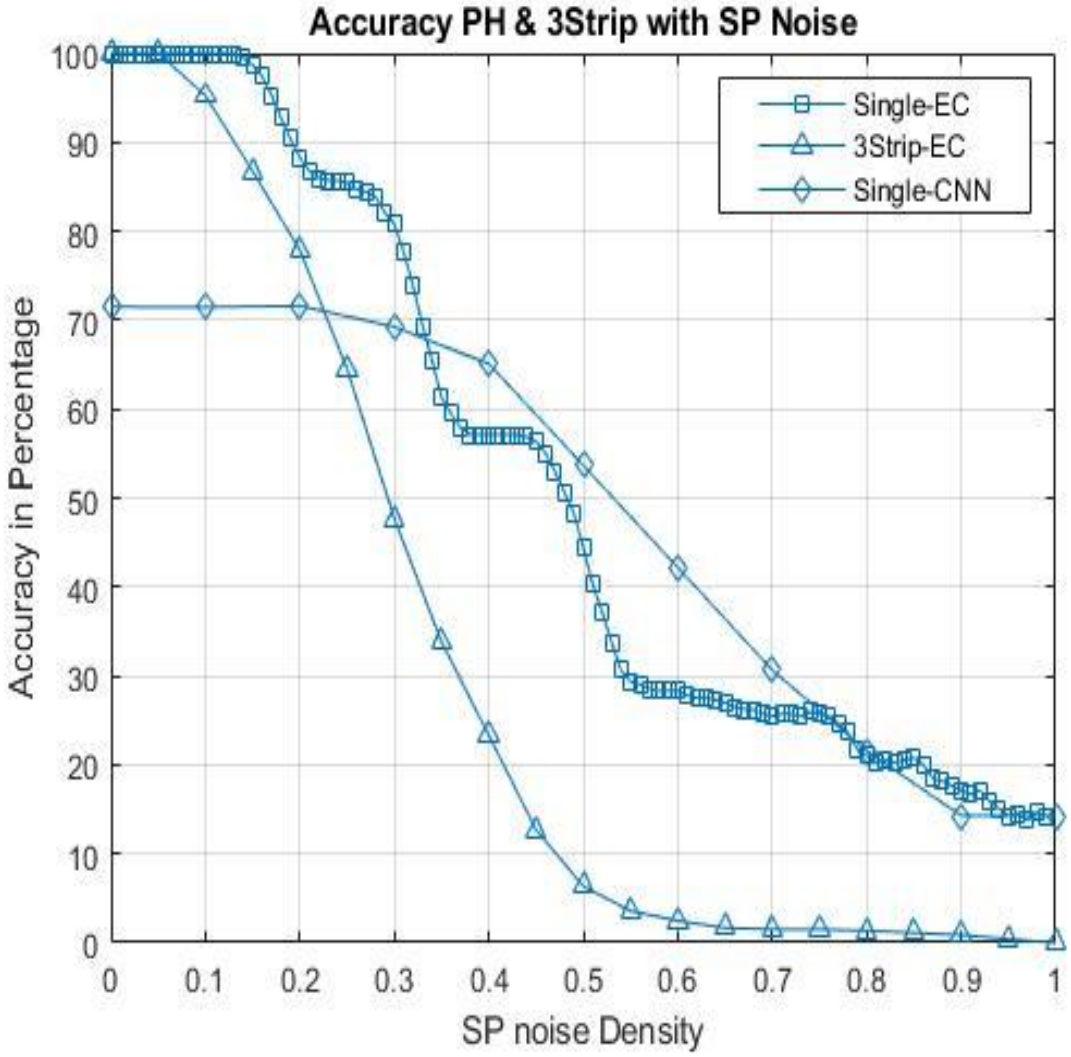


Figure 8: Accuracy Graph for PH & 3Strip with Salt & Pepper Noise

The following figure i.e. Figure 9 is used to represent accuracy vs noise graph. In this figure, the line with a square marker or Single-EC represents the accuracy for Protein segmented images with Gaussian noise using the Euclidean Distance method. The line with triangle marker or 3Strip-EC represents accuracy for whole strip images with Gaussian noise using the Euclidean Distance method. The line with a diamond marker or Single-CNN represents accuracy for Protein segmented images with Gaussian noise using CNN classifier. It is clear from the graph that accuracy has decreased as the variance of noise increases, in this graph initially at low-density noise the results of Euclidean distance comparatively better than CNN classifier but as noise increases Single CNN shows high accuracy as compared to the single-EC and 3strip-EC.

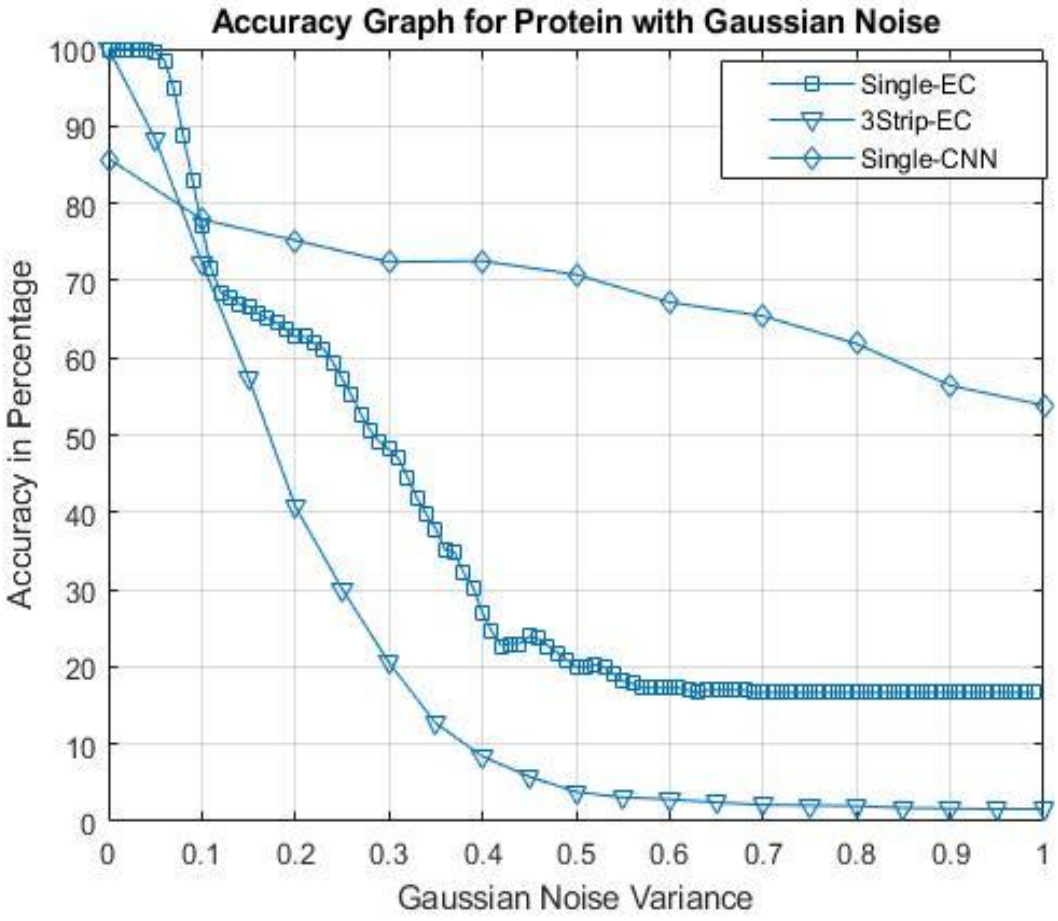


Figure 9: Accuracy Graph for Protein & 3Strip with Gaussian Noise

The following figure i.e. Figure 10 is used to represent accuracy vs noise graph. In this figure, the line with a square marker or Single-EC represents the accuracy for Protein segmented images with Salt & Pepper noise using the Euclidean Distance method. The line with triangle marker or 3Strip-EC represents accuracy for whole strip images with Salt &Pepper noise using the Euclidean Distance method. The line with a diamond marker or Single-CNN represents accuracy for Protein segmented images with Salt & Pepper noise using CNN classifier. It is clear from the graph that accuracy has decreased as the variance of noise increases. Single CNN shows high accuracy as compared to the single-EC and 3strip-EC.

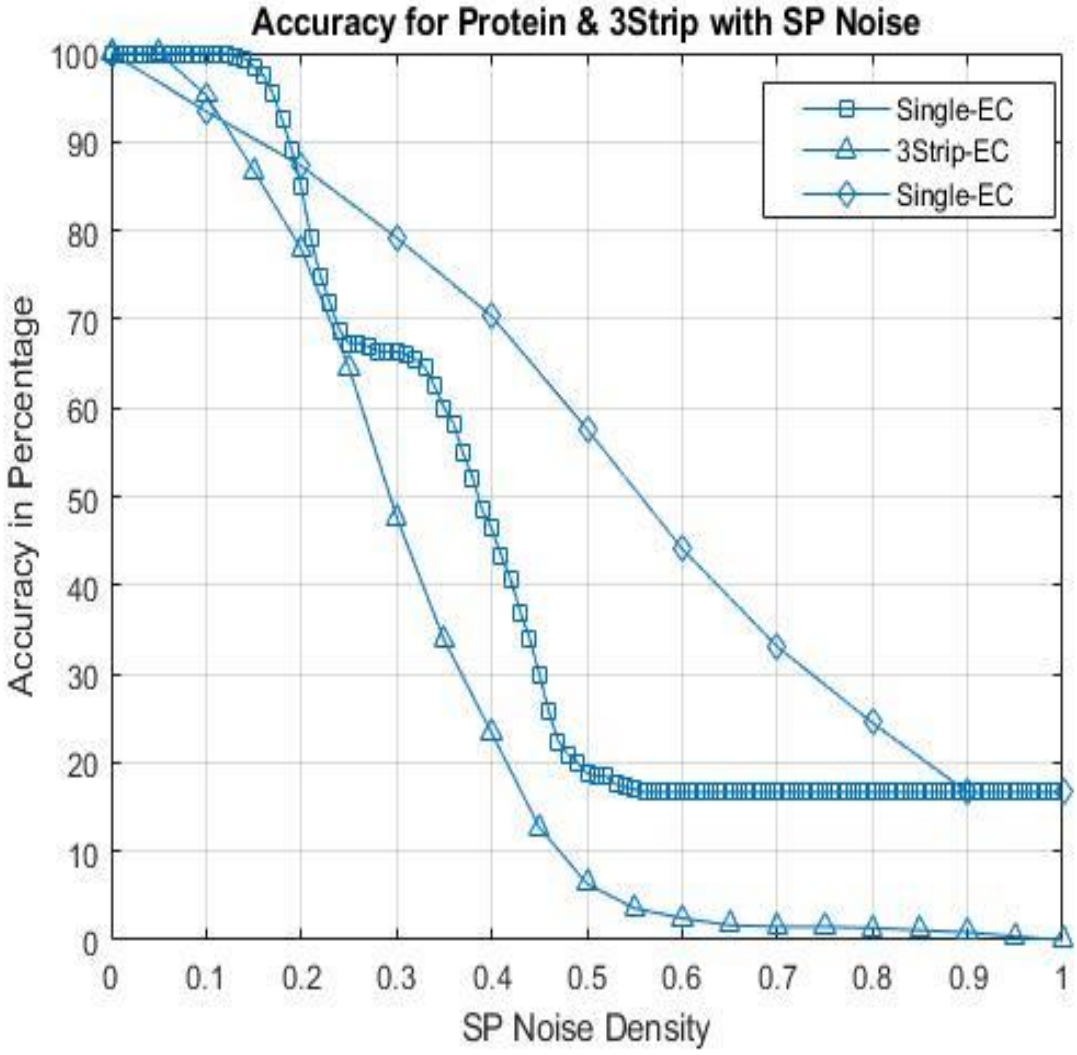


Figure 10: Accuracy Graph for Protein & 3Strip with Salt & Pepper Noise

5.3Analysis

Our proposed technique has been able to outclass the previously presented techniques in terms of accuracy. It has been able to achieve the highest accuracy because of the different experimentations carried out to analyze their results and hence present a better technique. The fact that experimentations have been performed to assess the methodology proposed in the previous chapter. Besides this, the results of our proposed methodology have also been evaluated and compared with the previously used method like Euclidean Distance. The main dataset used in this manuscript contains the urine test strip RGB images, which is developed using MATLAB.

Different noises i.e. Gaussian and salt & pepper have been applied on the dataset images to increase numbers. Initially, the dataset was comprised of three regent color strip images, which is used as an input for Euclidean Distance. After this, segmentation has been applied to the images and then the segmented images have been used for experimentation purposes in both methods. The accuracy vs noise variance graphs for the models has been presented in **Figure 5-10**. The experimentation results reveal that the regression model with the CNN classifier has a higher accuracy as compared with Euclidean Distance.

Different pre-processing techniques i.e. segmentation, data representation in the 2-d array, data normalization, and data reshaping has been applied before feeding our data o Convolutional Neural network. Validation is a very important part of the process to check the feasibility and accuracy of a technique. We have opted for a Simple K-Fold cross-validation technique that split our data into K parts for training, validation, and testing. To train and test our model, our data should be divided into three different datasets, consists of a training set, a validation set, and a test set.

5.4Summary

In this chapter, the results produced by the proposed technique have been presented in terms of accuracy. Different graphs have also been provided to provide the reader with a visual illustration of the performance of the algorithm. Moreover, the results that we attain from Euclidean Distance has also been presented in term of accuracy.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1Introduction

After the literature review, initially proposed methodology, experimentation, proposed methodology, and results and comparison, this manuscript has finally been concluded. In this chapter, an overview of the conducted research has been presented.

This thesis is dedicated to presenting a reliable machine learning for urinalysis. It aims to efficiently diagnose the stated problem, to prevent a fatality. For the stated purpose different experimentations have been conducted on the main dataset.

6.2Conclusion

Kidney disease, Urinary Tract Infection, Diabetes, and Tumor are the deadliest diseases and have been causing an increase in death rates all around the world. Many factors are involved in causing these fatal diseases among people of every gender, age, and profession. The physician mostly recommended urinalysis, one of the clinical diagnoses techniques to be the best diagnosis treatment. On the other hand, the recommended diagnosis technique is very costly and faces many challenges. Research has been conducted for the development of such techniques using machine learning which is not so much costly. With time, this disease becomes a fatal problem for public health all around the world. The reasons for this disease are a poor lifestyle, unawareness, and unhealthy consumption. Nowadays, it is a great challenge for the practitioners and hospital for a correct and accurate diagnosis. The recent development in computer technologies helps the health care system by facilitating in handling the collection of data for assisting in the decision making. In developed countries, the hospital manages the patient data in digital form.

If the fatal disease is not timely diagnosed, then it causes death even in days and in months also. For diagnosing purpose, urinalysis technique is mostly used but it is costly and causes side effects.

Moreover, numerous studies and research works have been carried out to diagnose these diseases and to a great extent, scientists have achieved success. To interpret the dataset, Machine Learning is used. Machine learning use computers for acquiring knowledge. ML using complex computation to deal with the health care center and emerged as disciplinary of science. Machine learning plays an important role in dealing with the healthcare dataset because the mostly dataset is in colossal form.

Numerous studies have been carried out to find out which ML techniques are best for the diagnosis of diseases related to urinalysis. Various studies show a comparison of different machine learning techniques for diagnostic procedures.

ML techniques can improve the results of Kidney disease, Urinary Tract Infection, Diabetes, and Tumor, but there is a requirement for improvement in the research for the logical thinking and selection of ML techniques for diagnosis purposes.

For the stated purpose urine strip images Dataset has been used. Our proposed technique has been able to outclass the previously presented techniques in terms of accuracy. It has been able to achieve the highest accuracy because of the different experimentations carried out to analyze their results and hence present a better technique. The fact that experimentations have been performed to assess the methodology proposed in the previous chapter. Besides this, the results of our proposed methodology have also been evaluated and compared with the previously used method like Euclidean Distance. The main dataset used in this manuscript contains the urine test strip RGB images, which is developed using MATLAB. Different noises i.e. Gaussian and salt & pepper have been applied on the dataset images to increase numbers. Initially, the dataset was comprised of three regent color strip images, which is used as an input for Euclidean Distance. After this, segmentation has been applied to the images and then the segmented images have been used for experimentation purposes. The accuracy vs noise variance graphs for the models

has been presented in **Figure 5-10**. The experimentation results reveal that the regression model with the CNN classifier has a higher accuracy as compared with Euclidean Distance.

Different pre-processing techniques i.e. segmentation, data representation in the 2-d array, data normalization, and data reshaping has been applied before feeding our data o Convolutional Neural network. Validation is a very important part of the process to check the feasibility and accuracy of a technique. We have opted for a Simple K-Fold cross-validation technique that split our data into K parts for training, validation, and testing. To test the proposed model, our data should be divided into three different categories, such as dataset for training, testing, and validation.

6.3Future Work

The accuracy of the proposed technique is promising. However, there is still a need for the betterment of results as far as accuracy is concerned. Currently where technology can be used to save valuable lives, why stop until profound perfection has been achieved. There is always a window available to ameliorate and this system can further be bettered.

When health is at stake, diseases need to be diagnosed timely to prevent them from threatening vital human life. That is the reason why these results can be improved, and hence further research is intended to be carried out to deliver better solutions to the challenges confronted by urinalysis. After improving the accuracy, for future work, an automated system for the diagnosis of urinalysis can be developed.

REFERENCES

[1] P. Wu, M. Zhu, P. Pu, and T. Jiang, "An Adaptive Sampling Ensemble Learning Method for Urinalysis Model", *2010 2nd International Conference on Information Engineering and Computer Science*, 2010. Available: 10.1109/iciecs.2010.5678258 [Accessed 21 July 2019].

[2] R. Ginardi, A. Saikhu, R. Sarno, D. Sunaryono, A. Kholimi and R. Shanty, "Intelligent Method for Dipstick Urinalysis Using Smartphone Camera", *Information and Communication Technology*, pp. 66-77, 2014. Available: 10.1007/978-3-642-55032-4_7 [Accessed 23 July 2019].

[3] F. de Boer et al., "Accurate and fast urinalysis in febrile patients by flow cytometry", *Infectious Diseases*, vol. 49, no. 5, pp. 380-387, 2017. Available: 10.1080/23744235.2016.1274048 [Accessed 26 July 2019].

[4] L. Egevad, B. Delahunt, J. Srigley and H. Samaratunga, "International Society of Urological Pathology (ISUP) grading of prostate cancer - An ISUP consensus on contemporary grading", *APMIS*, vol. 124, no. 6, pp. 433-435, 2016. Available: 10.1111/apm.12533 [Accessed 6 April 2020].

[5] L. Layfield and S. Frazier, "Quality assurance of anatomic pathology diagnoses: Comparison of alternate approaches", *Pathology - Research and Practice*, vol. 213, no. 2, pp. 126-129, 2017. Available: 10.1016/j.prp.2016.11.007 [Accessed 6 April 2020].

[6] E. Ranheim, "A novel program for clinical pathology training for residents emphasizing high-impact and attending-level learning opportunities", *Human Pathology*, vol. 45, no. 2, pp. 206-212, 2014. Available: 10.1016/j.humpath.2013.08.012 [Accessed 6 April 2020].

[7] M. Bluth, "Introduction: Molecular Medicine in the Common Era", *Clinics in Laboratory Medicine*, vol. 38, no. 2, pp. 209-213, 2018. Available: 10.1016/j.cll.2018.03.001 [Accessed 6 April 2020].

[8] N. Gabaj et al., "Urine test strip ascorbic acid interference: A risk analysis", *Clinica Chimica Acta*, vol. 493, p. S521, 2019. Available: 10.1016/j.cca.2019.03.1097 [Accessed 8 April 2020].

[9] J. McMillin, "Blood Glucose", *Ncbi.nlm.nih.gov*, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK248/. [Accessed: 08- Apr- 2020].

[10] M. Kim et al., "Presence of bilirubin in bronchobiliary fistula easily confirmed with urinary dipstick test and treated with embolization", *The Korean Journal of Internal Medicine*, vol. 32, no. 1, pp. 182-185, 2017. Available: 10.3904/kjim.2015.188 [Accessed 8 April 2020].

[11] D. Tidy, "Urine Ketones | Causes and Tests", *Patient.info,* 2020. [Online]. Available: https://patient.info/diabetes/urine-ketones. [Accessed: 08- Apr- 2020].

[12] "Urine specific gravity test: Procedure and results", *Medicalnewstoday.com,* 2020. [Online]. Available: https://www.medicalnewstoday.com/articles/322125#preparation. [Accessed: 08- Apr- 2020].

[13] E. MD, "Urinalysis", *Webpath.med.utah.edu*,2020. [Online]. Available: https://webpath.med.utah.edu/TUTORIAL/URINE/URINE.html. [Accessed: 08- Apr- 2020].

[14] P. Constable, A. Megahed and M. Hiew, "Measurement of urine pH and net acid excretion and their association with urine calcium excretion in periparturient dairy cows", *Journal of Dairy Science*, vol. 102, no. 12, pp. 11370-11383, 2019. Available: 10.3168/jds.2019-16805 [Accessed 8 April 2020].

[15] J. Robinson, A. Venner and I. Seiden-Long, "Urine protein detection by dipstick: No interference from alkalinity or specific gravity", *Clinical Biochemistry*, vol. 71, pp. 77-80, 2019. Available: 10.1016/j.clinbiochem.2019.07.005 [Accessed 8 April 2020].

[16] T. Kupka, L. Binder, D. Smith, B. Nelson, M. Wainscott and B. Glass, "Accuracy of urine urobilinogen and bilirubin assays in predicting liver function test abnormalities", *Annals of Emergency Medicine*, vol. 16, no. 11, pp. 1231-1235, 2019. Available: 10.1016/s0196-0644(87)80229-9 [Accessed 8 April 2020].

[17] [A. Fani, M. Rezaei, P. Samadian, P. Mirzajani, E. Vazifeh and A. Malekirad, "Correlation between nitrate in urine and daily diet among Arak-Iran citizens", *Toxicology Letters*, vol. 258, p. S174, 2016. Available: 10.1016/j.toxlet.2016.06.1650 [Accessed 8 April 2020].

[18] G. Rheinheimer, C. Bethoney, J. Boone, J. Fox and J. Stradinger, "Analytical sensitivity of urine dipstick leukocyte esterase tests", *Clinica Chimica Acta*, vol. 493, p. S663, 2019. Available: 10.1016/j.cca.2019.03.1397 [Accessed 8 April 2020].

[19] J. Wassan, H. Wang and H. Zheng, "Machine Learning in Bioinformatics", Encyclopedia of Bioinformatics and Computational Biology, vol. 1, pp. 300-308, 2018.

[20] M. Velikova, P. Lucas, R. Smeets and J. Terwisscha van Scheltinga, "Fully-automated interpretation of biochemical tests for decision support by smartphones", *2012 25th IEEE International Symposium on Computer-Based Medical Systems (CBMS)*, 2012. Available: 10.1109/cbms.2012.6266352 [Accessed 25 July 2019].

[21] C. Ongkum, K. Keawmitr and E. Boonchieng, "Analysis system for urine strip test using image processing technique", *2016 9th Biomedical Engineering International Conference (BMEiCON)*, 2016. Available: 10.1109/bmeicon.2016.7859610 [Accessed 22 July 2019].

[22] N. Frazzette, J. Dobson, A. Mukhtar and B. Burt, "Can we manufacture diagnostic test strips using an Inkjet printer?", *2015 IEEE Global Humanitarian Technology Conference (GHTC)*, 2015. Available: 10.1109/ghtc.2015.7344008 [Accessed 25 July 2019].

[23] J. Ro et al., "Development of a Plural Colorimeter Module for Urinalysis Strip Readers", *2007 International Conference on Convergence Information Technology (ICCIT 2007)*, 2007. Available: 10.1109/iccit.2007.411 [Accessed 25 July 2019].

[24] G. Gundermann, S. Sen, S. Ritter and K. Mehta, "Stamping: A low-cost manufacturing method to deposit assays", *2016 IEEE Global Humanitarian Technology Conference (GHTC)*, 2016. Available: 10.1109/ghtc.2016.7857341 [Accessed 25 July 2019].

[25] N. Arnett, A. Vergani, A. Winkler, S. Ritter and K. Mehta, "Inexpensive urinalysis test strips to screen for diabetes in developing countries", *2016 IEEE Global Humanitarian Technology Conference (GHTC)*, 2016. Available: 10.1109/ghtc.2016.7857339 [Accessed 22 July 2019].

[26] P. Wu, E. Goodman, T. Jiang and M. Pei, "A hybrid GA-based fuzzy classifying approach to urinary analysis modeling", *Proceedings of the 11th annual conference companion on Genetic and evolutionary computation conference - GECCO '09*, 2009. Available: 10.1145/1570256.1570381 [Accessed 27 July 2019].

[27] Y. Liang, R. Kang, C. Lian and Y. Mao, "An End-to-End System for Automatic Urinary Particle Recognition with Convolutional Neural Network", *Journal of Medical Systems*, vol. 42, no. 9, 2018. Available: 10.1007/s10916-018-1014-6 [Accessed 27 July 2019].

[28] R. Fletcher, N. Pignatelli, A. Jimenez-Galindo and S. Ghosh-Jerath, "Development of smart phone tools for printed diagnostics: Challenges and solutions", *2016 IEEE Global Humanitarian Technology Conference (GHTC)*, 2016. Available: 10.1109/ghtc.2016.7857355 [Accessed 22 July 2019].

[29] K. Konnaiyan, S. Cheemalapati, M. Gubanov and A. Pyayt, "mHealth Dipstick Analyzer for Monitoring of Pregnancy Complications", *IEEE Sensors Journal*, vol. 17, no. 22, pp. 7311-7316, 2017. Available: 10.1109/jsen.2017.2752722 [Accessed 23 July 2019].

[30] J. Neumeyer, J. Prince, A. Miller, B. Koeneman, S. Figueria and U. Kim, "Mobile urinalysis for maternal screening: Frugal medical screening solution and patient database to aid in prenatal healthcare for expecting mothers in the developing world", *2016 IEEE Global Humanitarian Technology Conference (GHTC)*, 2016. Available: 10.1109/ghtc.2016.7857337 [Accessed 23 July 2019].

[31] R. Sicat, G. Tangonan and M. Guico, "Patient-Centric Medical Database with Remote Urinalysis Test", *2009 WRI World Congress on Computer Science and Information Engineering*, 2009. Available: 10.1109/csie.2009.512 [Accessed 23 July 2019].

[32] D. Lee, W. Ick Jang and M. Yeon Jung, "A pocket-sized colorimetric urine reader for telemedicine in the developing countries", *2011 IEEE SENSORS Proceedings*, 2011. Available: 10.1109/icsens.2011.6127234 [Accessed 21 July 2019].

[33] M. Ra, M. Muhammad, C. Lim, S. Han, C. Jung and W. Kim, "Smartphone-Based Point-of-Care Urinalysis Under Variable Illumination", *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 6, pp. 1-11, 2018. Available: 10.1109/jtehm.2017.2765631 [Accessed 25 July 2019].

[34] G. Budianto, T. Harsono and H. Yuniarti, "Strip Test Analysis Using Image Processing for Diagnosing Diabetes and Kidney Stone Based on Smartphone", *2018 International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES-KCIC)*, 2018. Available: 10.1109/kcic.2018.8628569 [Accessed 25 July 2019].

[35] S. Mohamed, "UCHEK: an affordable smartphone based point of care diagnostic system for a low resource medical setup", *Appropriate Healthcare Technologies for Low Resource Settings (AHT 2014)*, 2014. Available: 10.1049/cp.2014.0783 [Accessed 25 July 2019].

[36] C. Florkowski, A. Don-Wauchope, N. Gimenez, K. Rodriguez-Capote, J. Wils and A. Zemlin, "Point-of-care testing (POCT) and evidence-based laboratory medicine (EBLM) – does it leverage any advantage in clinical decision making?", *Critical Reviews in Clinical Laboratory Sciences*, vol. 54, no. 7-8, pp. 471-494, 2017. Available: 10.1080/10408363.2017.1399336 [Accessed 27 July 2019].

[37] S. Ahn, K. Lee, D. Kim and V. Joseph, "Ubiquitous Healthcare Architecture Using SmartBidet and HomeServer with Embedded Urinalysis Agent", *Ubiquitous Intelligence and Computing*, pp. 205-213. Available: 10.1007/978-3-540-73549-6_21 [Accessed 27 July 2019].

[38] C. Naugler and D. Church, "Automation and artificial intelligence in the clinical laboratory", *Critical Reviews in Clinical Laboratory Sciences*, vol. 56, no. 2, pp. 98-110, 2019. Available: 10.1080/10408363.2018.1561640 [Accessed 26 July 2019].

[39] A. Bertao and T. Dong, "Stability of colorimetric results in the detection of urine biomarkers using a paper-based analytical device", *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2017. Available: 10.1109/embc.2017.8036793 [Accessed 23 July 2019].

[40] D. Kocer, F. Sarıguzel and C. Karakukcu, "Cutoff values for bacteria and leukocytes for urine sediment analyzer FUS200 in culture-positive urinary-tract infections", *Scandinavian Journal of Clinical and Laboratory Investigation*, vol. 74, no. 5, pp. 414-417, 2014. Available: 10.3109/00365513.2014.900189 [Accessed 26 July 2019].

[41] D. Bren-Cardali, L. Dao, J. Destruel, R. Fernandez, S. Figueira and U. Kim, "Urinalysis Screening for Rural Communities", *2018 IEEE Global Humanitarian Technology Conference (GHTC)*, 2018. Available: 10.1109/ghtc.2018.8601586 [Accessed 23 July 2019].

[42] M. Beňovská, O. Wiewiorka and J. Pinkavová, "Evaluation of FUS-2000 urine analyzer: analytical properties and particle recognition", *Scandinavian Journal of Clinical and Laboratory Investigation*, vol. 78, no. 1-2, pp. 143-148, 2018. Available: 10.1080/00365513.2017.1423108 [Accessed 26 July 2019].

[43] System for Diagnosing Disease Risk from Urine Tests", *Advances in Intelligent Systems and Computing*, pp. 241-248, 2018. Available: 10.1007/978-981-13-0344-9_20 [Accessed 27 July 2019].

[44] M. Fritzenwanker, C. Imirzalioglu, T. Chakraborty and F. Wagenlehner, "Modern diagnostic methods for urinary tract infections", *Expert Review of Anti-infective Therapy*, vol. 14, no. 11, pp. 1047-1063, 2016. Available: 10.1080/14787210.2016.1236685 [Accessed 26 July 2019].

[45] K. Shinozawa, N. Hagita, M. Furutani and R. Matsuoka, "A Data Mining Method for Finding Hidden Relationship in Blood and Urine Examination Items for Health Check", *Advances in*

Data Mining. Applications and Theoretical Aspects, pp. 44-50, 2009. Available: 10.1007/978-3-642-03067-3_5 [Accessed 27 July 2019].

[46] P. Andreini, S. Bonechi, M. Bianchini, A. Mecocci and V. Di Massa, "Automatic Image Classification for the Urinoculture Screening", *Intelligent Decision Technologies*, pp. 31-42, 2015. Available: 10.1007/978-3-319-19857-6_4 [Accessed 27 July 2019].

[47] J. Delanghe, "NEW SCREENING DIAGNOSTIC TECHNIQUES IN URINALYSIS", *Acta Clinica Belgica*, vol. 62, no. 3, pp. 155-161, 2007. Available: 10.1179/acb.2007.026 [Accessed 26 July 2019].

[48] J. Bekhof, B. Kollen, L. Groot-Jebbink, C. Deiman, S. van de Leur and H. van Straaten, "Validity and interobserver agreement of reagent strips for measurement of glucosuria", *Scandinavian Journal of Clinical and Laboratory Investigation*, vol. 71, no. 3, pp. 248-252, 2011. Available: 10.3109/00365513.2011.558109 [Accessed 27 July 2019].

[49] Y. Zhou and H. Zhou, "Automatic Classification and Recognition of Particles in Urinary Sediment Images", *Lecture Notes in Electrical Engineering*, pp. 1071-1078, 2011. Available: 10.1007/978-94-007-1839-5_116 [Accessed 27 July 2019].

[50] I. Duran, R. Leandro and J. Guevara-Coto, "Analysis of Different Pre-Processing Techniques to the Development of Machine Learning Predictors with Gene Expression Profiles", *2019 IV Jornadas Costarricenses de Investigación en Computación e Informática (JoCICI)*, 2019. Available: 10.1109/jocici48395.2019.9105145 [Accessed 12 June 2020].

[51] "Image Segmentation", *Mathworks.com*, 2020. [Online]. Available: https://www.mathworks.com/discovery/image-segmentation.html. [Accessed: 12- Jun- 2020].

[52] "Classify Images Using Convolutional Neural Networks & Python", *Medium*, 2020. [Online]. Available: https://medium.com/@randerson112358/classify-images-using-convolutional-neural-networks-python-a89cecc8c679. [Accessed: 13- Jun- 2020].

[53] P. Sane and R. Agrawal, "Pixel normalization from numeric data as input to neural networks: For machine learning and image processing", *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, 2017. Available: 10.1109/wispnet.2017.8300154 [Accessed 13 June 2020].

[54] S. Ghosh, N. Das and M. Nasipuri, "Reshaping inputs for convolutional neural network: Some common and uncommon methods", *Pattern Recognition*, vol. 93, pp. 79-94, 2019. Available: 10.1016/j.patcog.2019.04.009 [Accessed 14 June 2020].

[55] "5 Reasons why you should use Cross-Validation in your Data Science Projects", *Medium*, 2020. [Online]. Available: https://towardsdatascience.com/5-reasons-why-you-should-use-cross-validation-in-your-data-science-project-8163311a1e79. [Accessed: 15- Jun- 2020].

[56] "A Comprehensive Guide to Convolutional Neural Networks—the ELI5 way", *Medium*, 2020. [Online]. Available: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53. [Accessed: 18- Jun- 2020].

[57] A. Sharma V, "Understanding Activation Functions in Neural Networks", Medium, 2017. [Online]. Available: https://medium.com/the-theory-of-everything/understanding-activation-functions-in-neural-networks-9491262884e0. [Accessed: 20- Jun- 2020].

[58] T. Babs, "The Mathematics of Neural Networks – Coinmonks – Medium", Medium, 2018. [Online]. Available: https://medium.com/coinmonks/the-mathematics-of-neural-network-60a112dd3e05. [Accessed: 19- Jan- 2019].

[59] D. GUPTA, "Fundamentals of Deep Learning - Activation Functions and their use", Analytics Vidhya, 2017. [Online]. Available: https://www.analyticsvidhya.com/blog/2017/10/fundamentals-deep-learning-activation-functions-when-to-use-them/. [Accessed: 20- Jun- 2020].

[60] "Kulbear/deep-learning-nano-foundation", *GitHub*, 2020. [Online]. Available: https://github.com/Kulbear/deep-learning-nano-foundation/wiki/ReLU-and-Softmax-Activation-Functions. [Accessed: 20- Jun- 2020].

[61] S. Ruder, "An overview of gradient descent optimization algorithms", Sebastian Ruder, 2016. [Online]. Available: http://ruder.io/optimizing-gradient-descent/index.html#adam. [Accessed: 20- Jun- 2020].

[62] "Mean squared error loss function | Peltarion Platform", *Peltarion.com*, 2020. [Online]. Available: https://peltarion.com/knowledge-center/documentation/modeling-view/build-an-ai-model/loss-functions/mean-squared-error. [Accessed: 22- Jun- 2020].

[63] "Euclidean Distance - an overview | ScienceDirect Topics", *Sciencedirect.com*, 2020. [Online]. Available: https://www.sciencedirect.com/topics/mathematics/euclidean-distance. [Accessed: 22- Jun- 2020].

[64] C. Tseng and S. Lee, "A Low-Light Color Image Enhancement Method on CIELAB Space", *2018 IEEE 7th Global Conference on Consumer Electronics (GCCE)*, 2018. Available: 10.1109/gcce.2018.8574809 [Accessed 22 June 2020].

[65] Y. Khan, U. Qamar, N. Yousaf and A. Khan, "Machine Learning Techniques for Heart Disease Datasets", *Proceedings of the 2019 11th International Conference on Machine Learning and Computing - ICMLC '19*, 2019. Available: 10.1145/3318299.3318343 [Accessed 25 June 2020].